THE INFLUENCE OF GLOTTAL EXCITATION FUNCTIONS ON THE QUALITY OF SYNTHETIC SPEECH

BY

JING-JONG YEA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1983

ACKNOWLEDGEMENTS

I would like to express my gratitude to my adviser, Dr. Donald G. Childers, whose constant encouragement and advice have made the completion of this dissertation possible. I would also thank Dr. G. P. Moore and Dr. D. Hicks for their valuable suggestions in designing the listening test and evaluating the quality of synthetic speech. Thanks are due to Mr. Ralph Haskew for his excellent job in typing and to my fellow students in the Mind-Machine Interaction Laboratory for their help in many ways.

Finally, I would like to thank my parents for motivating my curiosity for knowledge and my wife for her love, encouragement, and support.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

THE INFLUENCE OF GLOTTAL EXCITATION FUNCTIONS ON THE QUALITY OF SYNTHETIC SPEECH

By

Jing-Jong Yea

August 1983

Chairman: Donald G. Childers
Major Department: Electrical Engineering

One of the factors affecting the quality of synthetic speech is the improper modeling of the glottal excitation function. This includes the use of an impulse or a stylized waveform as the glottal excitation, and neglecting the effect of source-tract interaction.

The results of our research improve the quality of synthetic speech by using a formant synthesis scheme which adopts a more accurate model of the glottal excitation. This is accomplished by using a "glottal area" excitation function. In this scheme, we use the glottal area function to control the time-varying glottal impedance in an equivalent circuit of the vocal system. The output of the circuit is a time-varying "glottal volume velocity" function which includes the effect of source-tract interaction. This glottal volume velocity function is then used as the glottal excitation of the formant synthesizer. We used this technique to synthesize sentences which sounded as if they were spoken repeatedly by a male, a female, and a child. We used two other glottal excitation functions as

well to synthesize the same sentences for comparison purposes. The vocal tract transfer functions are kept the same regardless of the forms of the glottal excitation. The only difference is that one excitation waveform is an impulse and the other a stylized waveform.

The sentences generated by the three glottal excitation functions are compared in a formal listening test using the natural sentence as the reference signal. The result of the listening test shows that the proposed glottal excitation function can produce more natural sounding speech than the other two excitation waveforms. Thus, the results of our study suggest that source-tract interaction is an important factor for synthesizing high quality speech.

CHAPTER 1
INTRODUCTION

Speech synthesis has a long and interesting history. The first speech synthesizer was built by von Kempelen in 1791 [1]. His speech synthesizer was mechanical in nature and capable of producing only a few vowels. At that time the interest in speech synthesis was academic or for entertainment purposes, which remained the case until the turn of this century.

The modern speech synthesizer is electrical in nature. The evolution of electronic technology coupled with the widespread use of computers caused the interest in speech synthesis to assume a broader basis. There are three types of applications for speech synthesis.

The first is academic interest in the physiology and acoustics of speech production. The speech synthesizer is now a standard research tool in phonetics and speech perception experiments because of the convenience of controlling the parameters and reproducing the results.

The second concerns efficient coding of speech information for communication at a distance. In this application, the speech synthesizer is a part of the speech analysis/synthesis system which reconstructs speech from a set of parameters (Figure 1). In this way, the transmission bit rate (bandwidth) can be greatly reduced (by about an order of 10).

The third type of application is in computer voice response. Because of the widespread use of computers, there is a need for a more natural means

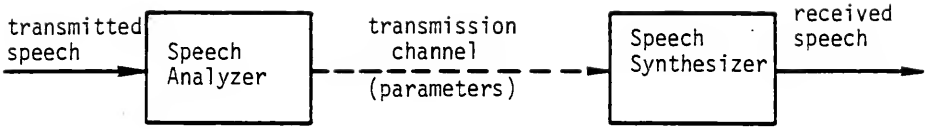


Figure 1. Speech Synthesizer as a part of communication system.

of communication between human and machine. Speech synthesizers enable the computer to communicate in terms of human speech. In this case, a set of rules are stored in the computer memory to generate parameters which in turn control a speech synthesizer to generate speech (Figure 2).

One common characteristic of speech synthesizers is that it requires only a small number of parameters to generate a large number of speech samples. The price we have to pay for this data compression is that the quality of synthetic speech is generally not comparable to that of human speech. The quality of synthetic speech is generally considered to consist of two factors:

(1) Intelligibility--does the synthetic speech convey the intended message correctly?

(2) Naturalness--does the synthesized speech sound like human generated speech?

Many previous papers have been concerned with improving the intelligibility of synthetic speech. The result is that synthetic speech is highly intelligible but less natural. In this research, we are concerned with the problem of improving the naturalness of synthetic speech. We are particularly interested in investigating the effect of the glottal source waveform on the quality of speech. Before we go into the details of the research problem, let us first discuss the general background of speech synthesis.

Overview of Speech Synthesis

The term "speech synthesis" has been used vaguely in the literature as "a procedure to produce or reproduce speech from some representation of speech." In this sense, speech reproduction from such waveform coding schemes as Pulse Code Modulation (PCM) and Delta Modulation (DM) can also be viewed as a speech synthesis process. In this study, however, we shall

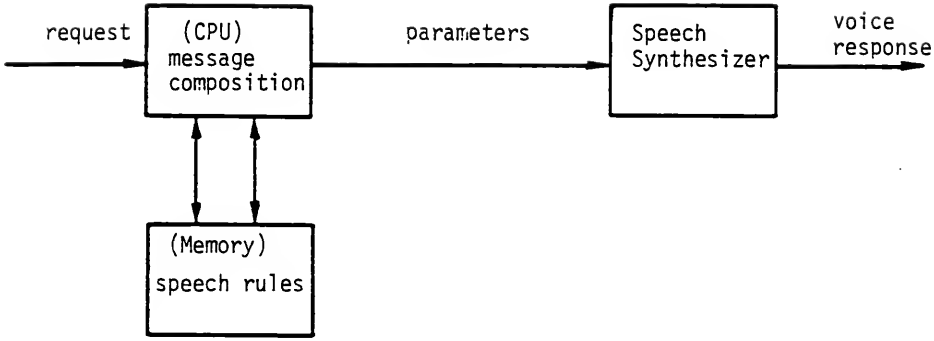


Figure 2. Speech Synthesizer as a part of the computer voice response system.

adopt a more restricted view of speech synthesis. We will define the term "speech synthesis" as "the procedure to produce speech from a parametric model of speech." Under this definition, the speech synthesis schemes can be classified in terms of their underlying model. Modern speech synthesizers fall into three basic categories: articulatory synthesis, formant synthesis, and linear prediction (LP) synthesis.

Articulatory synthesis is based on a physiological model of speech production [1,2,3]. As shown in Figure 3, this model attempts to simulate the mechanical motions of the articulators. The model then controls an equivalent circuit of the vocal system (Figure 4) to produce speech. The articulatory study is very useful for phonetic studies because it directly relates the articulator movements with the resultant acoustic events. The problems with articulatory synthesis are that a) knowledge about the articulator positions is not easily available for continuous speech [4], and b) the computation is more complex than for the other two synthesis schemes. These two difficulties have deferred the practical application of the articulatory synthesis in communication and computer voice response.

Formant synthesis is based on an acoustic model of speech production. This approach capitalizes on the fact that the formants, or vocal tract resonances, are the most important factors in deciding the content of speech. The formant synthesizer consists of two parts (Figure 5): 1) a set of resonance circuits which decides the speech content, and 2) an excitation source which decides the voicing information. The resonant circuits can be connected in serial (cascade formant synthesizer) or in parallel (parallel formant synthesizer).

Both the formant and the voicing information can be directly measured from the speech signal by digital speech processing techniques. This makes

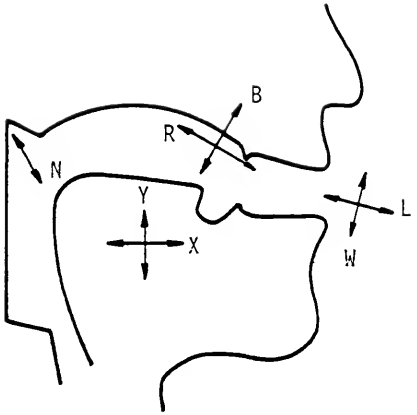


Figure 3. Articulatory model of the human vocal tract.

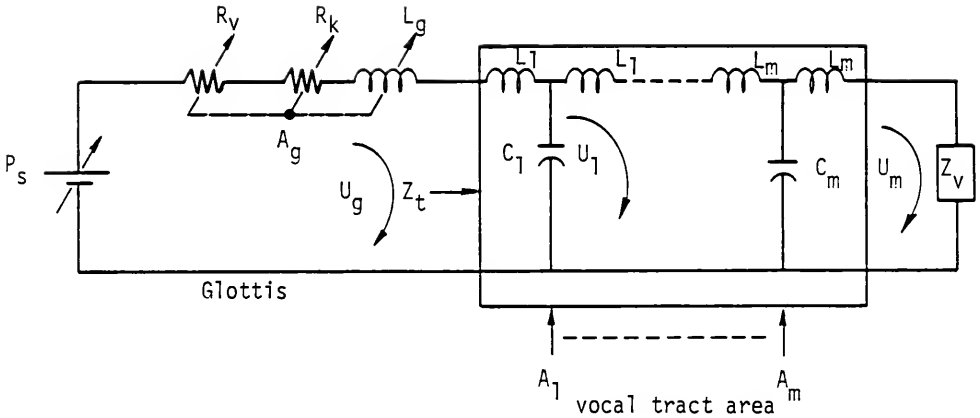


Figure 4. Equivalent circuit for the vocal system.

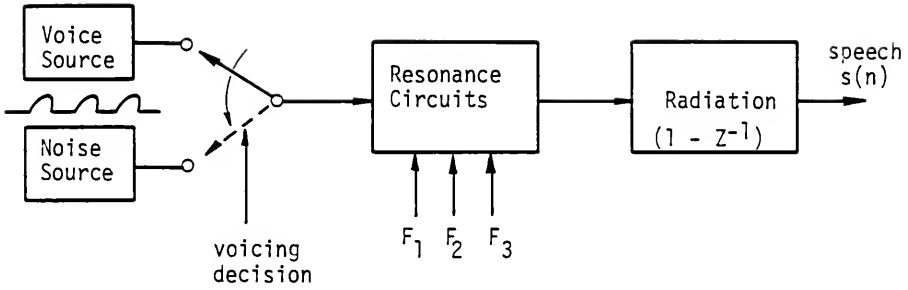


Figure 5. Block diagram of formant synthesizer.

the formant synthesis approach suitable for applications in communication and computer voice response.

Linear prediction (LP) synthesis is based on a mathematical model of the speech signal. The basic assumption of this model is that the speech signal is the output of a linear time invariant recursive filter (Figure 6). This is equivalent to predicting the current sample of speech by a linear combination of previous samples, hence the name linear prediction. It has been shown that this model is exact for voiced speech [5], but only an approximation for unvoiced and nasal sound. Linear prediction analysis/synthesis is becoming one of the most popular speech processing techniques because the linear prediction coefficients (a_k) can be obtained from the speech signal by a very efficient algorithm. Thus, linear prediction analysis/synthesis is finding many applications in different areas of speech research. The disadvantage is that unlike formant or articulatory synthesis, there is no direct physical correspondence for the linear prediction coefficients.

We have briefly discussed the three dominant speech synthesis strategies with their advantages and disadvantages. The choice of a particular speech synthesis strategy depends on the application on hand. In studying the physiology and acoustics of speech production, the formant synthesis or the articulatory synthesis should be used because they relate the speech to the physical parameters directly. For communication applications, the LP synthesis or formant synthesis should be used because of the ease of obtaining the parameters from the speech signal. The LP analysis is particularly suitable for this purpose because the analysis and synthesis can be implemented in real time in this case. For computer voice response, all three are suitable depending on the particular application with the LP and formant synthesis techniques being more suitable for the limited vocabulary case,

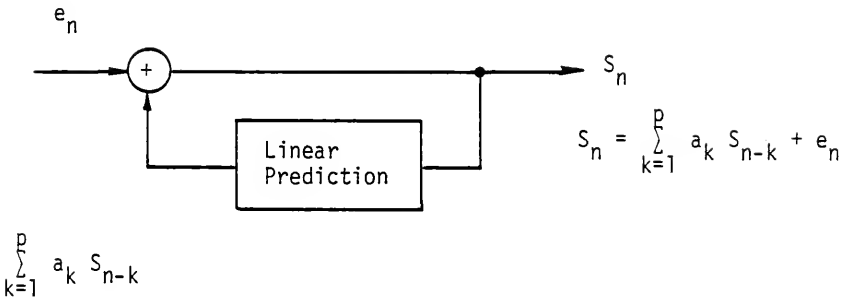


Figure 6. Block diagram of the linear prediction model of speech.

while the formant and articulatory synthesis techniques are more suitable for unlimited text-to-speech synthesis.

In this study, we are interested in studying the effect of glottal excitation on the quality of synthetic speech. Since the linear prediction model is a mathematical representation, the concept of glottal excitation is meaningless. This leaves us the choice of formant or articulatory synthesis. In both of these two synthesis schemes, the glottal excitation function is an explicit part of the synthesizer. Since formant synthesis has the advantages of 1) computational efficiency and 2) ease of obtaining the synthesis parameters directly from speech, we decided to use the formant synthesis scheme in our study.

Research Problem

The purpose of any speech synthesis study is usually to improve the quality of synthetic speech. It includes the improvement of intelligibility and naturalness.

To date, hundreds of speech synthesis schemes have been proposed, but few if any of them are able to generate natural sounding synthetic speech.

This difficulty in synthesizing natural sounding speech is caused mainly by inadequacies in the modeling of the human vocal system. The ways in which formant models usually differ from the human vocal system include the following [6]:

(a) Nasalized vowels are either not treated differently from nonnasalized vowels, or are generated using one additional pole-zero pair added to the transfer function without any corresponding change of formant bandwidth due to additional damping.

(b) No attempt is made to copy the natural glottal excitation waveform for voiced speech. Usually a pulse with minimum-phase spectrum

shaping is used. Otherwise, a stylized glottal pulse is used that approximates well to typical natural shapes in general features, but not in details.

(c) The modification of formant frequencies and bandwidths by source-tract interaction is ignored.

(d) Many consonant sounds are dealt with by special arrangements that do not closely approximate the acoustic production system.

(e) Bandwidth limitations on control signals prevent very rapid changes of amplitude (such as occur on stop burst), or formant frequency (such as at certain consonant-to-vowel boundaries).

(f) Mixed excitation for voiced fricatives and stops is often not provided.

Items a, d, e, and f are related to the intelligibility problem of speech synthesis, while items b and c are concerned with the naturalness of synthesis. Since we are mainly interested in the naturalness of synthetic speech, we shall compare previous studies which addressed items b and c.

Rosenberg [7] first reported a study comparing the effect of different glottal source waveforms on the naturalness of synthetic speech. He used a pitch synchronous pole-zero analysis technique [8] to extract formant frequencies, bandwidths and glottal volume velocity waveforms. He resynthesized a set of utterances (words) using natural glottal volume velocity and some idealized pulses. The results of his study showed that the listeners preferred a particular idealized pulse to the real glottal waveform.

Holmes [6] used a parallel formant synthesizer to synthesize sentences. His goal was to generate speech indistinguishable subjectively from the human generated speech. The glottal waveforms he used included the real

glottal waveform and idealized pulses suggested by Rosenberg. He also incorporated the source-tract interaction by adjusting the formant bandwidth during the glottal opening period. The results of an informal listening test showed that 60% of his subjects preferred the natural glottal waveform over the idealized waveform under the most critical listening conditions, i.e., using earphones.

There is a contradiction between the above two studies. Rosenberg's study showed that synthetic speech generated by an idealized waveform sounds better than the synthetic speech generated by the real glottal volume velocity. Holmes' study, however, concluded the opposite. We have carefully studied the above two papers and found that there are some differences which may contribute to the contradiction.

(a) Rosenberg used a serial formant synthesizer while Holmes used a parallel formant synthesizer. The comparison of the serial and parallel formant synthesizers has been studied before [6]. In general, the serial formant synthesizer is capable of generating a natural spectrum envelope for vowels without any additional controls. The disadvantage of the serial formant synthesizer is that it is not very good in synthesizing consonants. The parallel formant synthesizer, on the other hand, needs additional amplitude control for each formant so that the spectrum envelope matches that of natural speech. The advantage is that the parallel formant synthesizer is suitable for generating both vowels and consonants.

(b) A more important distinction between the two studies is the assumption about the dependency between the glottal source and the vocal tract. Rosenberg assumed that the source and tract are independent of each other and so the glottal waveforms should be similar for all vowels. He then used the same glottal volume velocity for

synthesizing different words. This procedure was known to introduce undesired quality to the synthetic speech [9]. Holmes, on the other hand, incorporated source-tract interaction in the synthesis by manually adjusting the formant bandwidths.

The above arguments seem to be in favor of Holmes' conclusion that the real glottal volume velocity waveform is important for the generation of natural sounding speech. Also, they indicate that the source-tract interaction is a significant factor in deciding the quality of synthetic speech [10].

Other studies [11,12] investigated the usefulness of the glottal area function as the excitation for speech synthesizers. The glottal area function is a measure of the opening of the glottis during voiced speech. It is obtained from ultra high speed films of the glottis. The formant data were extracted from digitalized speech using the linear prediction analysis. The speech (vowel) was re-synthesized using a volume velocity waveform which was derived from the glottal area function without taking account of the source-tract interaction. The results of this study are

(a) The synthetic vowel and its natural counterpart were distinguishable.

(b) The synthetic vowel did not sound natural.

(c) The real speech also suffered from a lack of naturalness. This could be attributed to the abnormal condition under which the speech was recorded (with the laryngeal mirror inserted into the mouth).

According to the author, three important factors for the unnaturalness of the synthetic speech were the interpolation of the glottal area (which was required due to low camera speed); the absence of source-tract interaction; and the lack of higher order formants, which contribute perceptually to naturalness.

From the above discussion, it is clear that the source-tract interaction is an important factor for the generation of natural sounding speech. Recently, there were a series of papers on the analysis of the glottal source excitation and the source-tract interaction [13,14]. They all agreed that the source-tract interaction affects the voice quality. In particular, they studied the skewing of the glottal volume velocity waveform due to the source-tract interaction and related it to the quality of voice. Another aspect of source-tract interaction, such as the formant ripples found in the inverse-filtered waveform [15], also influences the quality of speech. We will refer to this latter aspect as the source-tract interaction later in our discussion.

From the above discussion, we concluded that the glottal source waveform and the source-tract interaction are both important to the generation of natural sounding speech. Thus, it is important that we include these features in the glottal source of a speech synthesizer.

Proposed Research

In this research, we propose to study the influence of the glottal source function on the quality of synthetic speech. We do this by comparing three types of glottal excitation functions:

(a) Impulse excitation with glottal shaping is an excitation source which approximates the spectral characteristics of a real glottal source excitation by a proper glottal shaping filter. This excitation is similar to the glottal excitation for the LP synthesizer.

(b) Idealized glottal volume velocity waveform was first proposed by Rosenberg [7] and later used by Holmes [6] and Fant [13] in their studies. This excitation function approximates both the spectral

and time-domain characteristics of the real glottal excitation. We will discuss this excitation source in Chapter 3.

(c) The glottal area excitation source, unlike the glottal area excitation used in the Nadal-Suris study [11], includes the effect of source-tract interaction. This is done by transforming the glottal area function into glottal volume velocity using an equivalent circuit to the vocal system [15]. Thus, besides approximating the general spectral and time-domain characteristics of the glottal source, this excitation also contains the details of source-tract interaction.

The three types of glottal excitations are summarized in Figure 7. The main difference between the first two glottal excitations is that the first one only approximates the spectral characteristics of the glottal source (i.e., a -12 dB/octave drop in the spectral envelope), while the second excitation approximates both the spectral and time-domain characteristics (triangular-like waveshape). The difference between the second and third excitation function is that the second does not include the source-tract interaction (formant ripples) while the third one does.

Except for the excitation functions, the formant synthesizer is similar to Klatt's formant synthesizer [16]. We will use these three types of excitation functions to synthesize sentences. Three sentences will be synthesized. The synthesis parameters are extracted from sentences produced by an adult male, an adult female, and a female child, respectively. In this way, a wide range of pitch variations are covered so we can look into the influence of pitch on the quality of synthesis.

The human generated sentences will be used as reference sentences in a formal listening test to evaluate the quality of synthetic sentences.

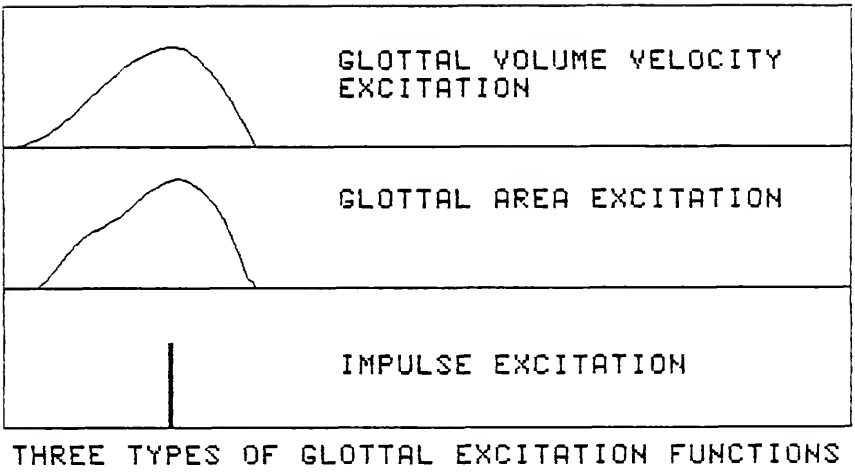


Figure 7. Three types of glottal excitation functions used in this study.

The result of this study will contribute to our understanding of the glottal source function. In particular, we will be able to learn the relative importance of spectral characteristics, time-domain characteristics, and source-tract interaction on the quality of speech. The study will also contribute to our knowledge about the difference between male/female and adult/child voices. Such knowledge will be useful for the construction of speech synthesizers which are capable of generating natural sounding speech.

CHAPTER 2
SPEECH ANALYSIS

As we defined in Chapter 1, speech synthesis is a procedure to produce speech from a parametric model. Thus, before synthesizing speech, we have to obtain the parameters of the model. The process of obtaining the parameters from speech is called "speech analysis." Speech analysis is the basis of all speech research. In this chapter we will discuss the analysis of speech to obtain parameters necessary for formant synthesis. These parameters include the pitch/voicing information, the formant frequencies and bandwidths, and the glottal source waveform. Before we go into the details of speech analysis, however, let us discuss the acoustic model of speech production.

Model of Speech Production

Formant synthesis is based on the acoustic model of speech production [17]. In this model, the speech signal is the response of the vocal tract filter system to one or more sound sources. The source-filter theory of speech production is exemplified by the equivalent circuit in Figure 8 in which the voltage is the equivalent of the acoustic sound pressure and the current is the equivalent of the volume velocity.

Sound Sources

There are two types of sound sources: the glottal sound source and the fricative sound source. For the glottal sound source, E_s is the lung pressure

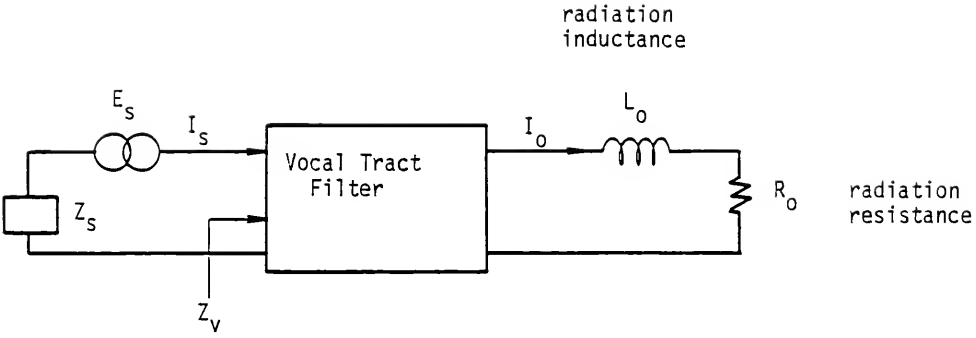


Figure 8. Equivalent circuit of speech production.

which pushes air through the vocal folds to make them vibrate and thus create a modulated air flow. For the fricative sound, E_s represents the turbulent air pressure generated by pushing air through the vocal tract constriction. The term Z_s represents the internal impedance of the source and the impedance of the cavities behind the source. The vocal tract input impedance is represented by Z_v. In the linear model of speech production it is assumed that Z_v is negligible compared to Z_s, so the source volume velocity is determined by E_s and Z_s only; i.e., the source and the vocal tract can be treated separately. However, there is evidence from inverse filtering experiments [18] that the glottal source volume velocity contains small formant ripples due to the loading of the vocal tract input impedance. This loading effect is referred to as the "source-tract interaction" in the literature. It is believed that the source-tract interaction will affect the quality of synthetic speech.

Since we are interested in synthesizing high quality speech, it is important that we include the source-tract interaction in the glottal source model.

Besides the source-tract interaction, the sound sources are characterized by two parameters, voicing and pitch. The voicing parameter decides whether the sound is voiced or fricative. The pitch, or fundamental frequency parameter, decides the vibration period of the glottal sound source. Both pitch and voicing can be determined from the speech signal as we will see in the next section.

The Vocal Tract Filter

The vocal tract filter models the transmission between the source volume velocity, I_s, and the mouth (or nose for nasal sound) volume velocity, I_o. Since the human vocal tract cannot change too rapidly, the vocal tract

filter can be viewed as a linear time-invariant filter for a short period of time (20 ms). In the production of vowels, the transmission characteristics can be modeled as a cascade of resonant circuits (corresponding to peaks in frequency response), referred to as formants in the literature. In the production of nasals and fricatives, the transmission characteristics have to be modeled as a combination of resonances and antiresonances (peaks and valleys, respectively, in the frequency response), called formants and antiformants.

Formants are the most important parameters of the vocal tract filter because the human ears are more sensitive to peaks in the sound spectrum [19]. Each formant is characterized by its frequency and bandwidth. For the speech signal sampled at 10 KHz, the vocal tract filter can be represented by five formants for adult males, or four formants for adult females and children. We will discuss the algorithm for extracting formants in the third section of this chapter.

Antiformants are very important for synthesizing nasalized sounds. However, there are no effective methods for estimating the antiformants. We have developed a method for detecting the existence of nasal sounds which is useful in our speech synthesis study. This algorithm can also be used as an intermediate step for estimating the antiformant frequency. We will present this algorithm later in Appendix B.

Radiation Effect

The transformation between the mouth volume velocity and sound pressure at the listener's ear is called the radiation effect. This effect can be approximated by a first order differentiation [19]. In the circuit of Figure 8, the radiation effect is represented by a simple R-L circuit.

Pitch and Voicing Analysis

Pitch and voicing analysis is one of the most important problems in speech processing. Because of its importance, many solutions to this problem have been proposed. All of the proposed schemes have their limitations, and it is safe to say that no presently available pitch detection scheme can be expected to give perfectly satisfactory results across a wide range of speakers, applications, and operating environments. The difficulty of pitch/voicing detection (or for speech analysis as a whole) stems from the fact that the speech signal is the convolution of the source waveform and the filter response. The effect of the tract will then interfere with the process of pitch/voicing determination, and vice-versa. Thus, most pitch detection algorithms first attempt to separate the source and the filter characteristics. We have implemented two algorithms for pitch/voicing detection: the cepstrum method, and the modified autocorrelation method. We will briefly discuss the principles of these two algorithms and then compare their performance in terms of real speech data.

The Cepstrum Pitch Detection Method

This approach is accomplished in the frequency domain. The cepstrum, defined as the power spectrum of the logarithm of the power spectrum, has a peak corresponding to the pitch period of the voiced-speech segment being analyzed [20]. Also, by detecting the presence or absence of the peak in the expected pitch dynamic range, we can decide whether the speech segment is voiced or unvoiced.

The way in which the cepstrum method separates the source and the filter characteristics is explained as follows. Since the voiced speech is the response of the vocal tract filter to the glottal volume velocity, we

can express the voiced speech as

$$s(t) = h(t) * u(t) \tag{1}$$

or

$$S(\omega) = H(\omega) \cdot U(\omega) \tag{2}$$

where $h(t)$ is the vocal tract impulse response,

$H(\omega)$ is the Fourier transform of $h(t)$,

$u(t)$ is the glottal volume velocity, and

$U(\omega)$ is the Fourier transform of $u(t)$.

The power spectrum of speech is the magnitude square of the Fourier transform, and can be expressed as

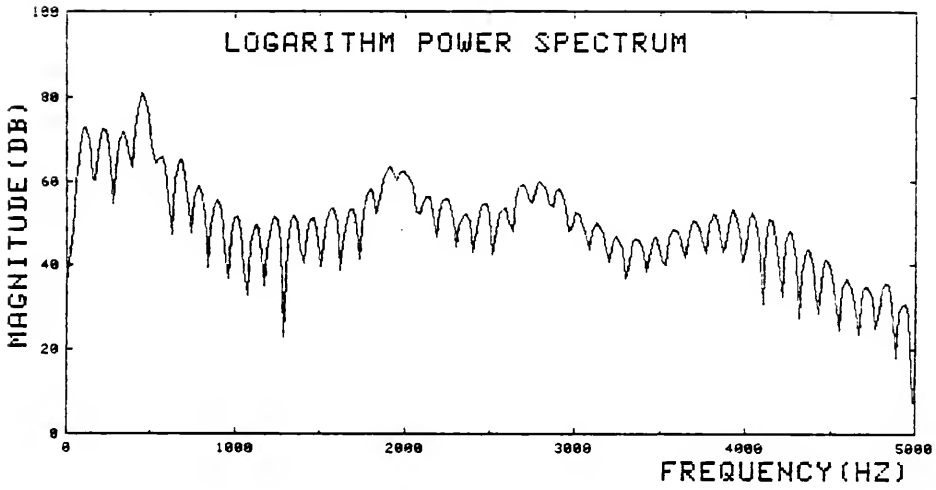
$$\left|S(\omega)\right|^2 = \left|H(\omega)\right|^2 \cdot \left|U(\omega)\right|^2 \tag{3}$$

One simple way to separate the contribution of the source and the tract is to take the logarithm of $\left|S(\omega)\right|^2$, thereby changing the multiplication into addition. Next take the Fourier transform which gives

$$\log \left|S(\omega)\right|^2 = \log \left|H(\omega)\right|^2 + \log \left|U(\omega)\right|^2 \tag{4}$$

$$F[\log \left|S(\omega)\right|^2] = F[\log \left|H(\omega)\right|^2] + F[\log \left|U(\omega)\right|^2] \tag{5}$$

The source and tract effects are now additive rather than multiplicative. The importance of this can be explained with the assistance of Figure 9. The effect of the vocal tract is to produce a "low frequency" spectral envelope in the logarithm spectrum, while the periodicity of the glottal source manifests itself as "high frequency" ripples in the logarithm spectrum. Therefore, the spectrum of the logarithm power spectrum has a sharp peak corresponding to the high frequency source ripples in the logarithm power spectrum and a broader peak corresponding to the low-frequency formant



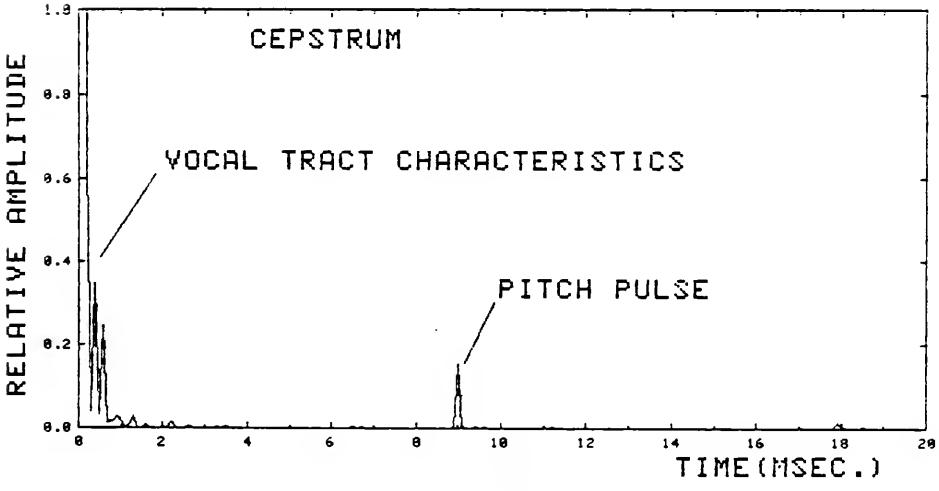


Figure 9. Logarithm power spectrum (top) of a voiced speech segment showing a spectral periodicity resulting from the pitch periodicity of speech. The power spectrum of the logarithm spectrum, or cepstrum (bottom), therefore has a sharp peak corresponding to this spectral periodicity.

structure in the logarithm spectrum. We can make the peak corresponding to the source periodicity more pronounced by taking a square, and hence obtain the cepstrum.

The cepstrum of speech can be computed on a general purpose computer using the Fast Fourier Transform (FFT) algorithm. The location and amplitude of the peak can then be determined, and an algorithm can be used to decide whether it corresponds to the pitch period or not.

The cepstrum pitch detection algorithm has been implemented on a Data General NOVA 4 computer. The program is called "PITCH." Appendix A has a sample dialog of using this program. We will compare the performance of the cepstrum method with the modified autocorrelation method later in this section.

The Modified Autocorrelation Method

The modified autocorrelation method is a time domain approach for pitch detection [21]. This method first removes the formant structure (vocal tract characteristics) by a technique called center clipping. The autocorrelation of the center clipped speech is then used to determine the pitch period and voicing information.

The center clipped speech signal is obtained in the manner illustrated in Figure 10. A segment of speech to be used in pitch detection is shown in the upper plot. For this segment, the maximum amplitude, A_{max}, is found and the clipping level, C_L, is set equal to a fixed percentage of A_{max}. From Figure 10 it can be seen that for samples above C_L, the output of the center clipper is equal to the input minus the clipping level. For samples below the clipping level the output is zero. The center clipped speech is shown in the lower plot of Figure 10. It can be seen that the formant

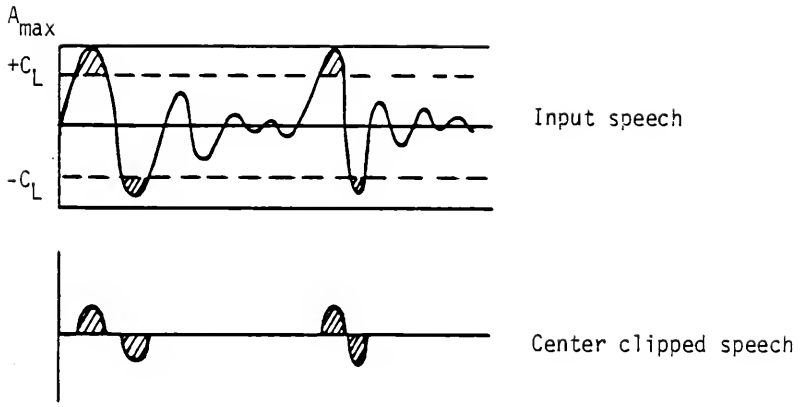


Figure 10. Illustration of the process of center clipping.

oscillation is removed from the center clipped speech. Thus the autocorrelation function is free from the interference of the formant oscillations. This is very important for the purpose of pitch detection. Because formant oscillations manifest themselves as peaks in the autocorrelation function and sometimes these peaks are greater than the peak due to the pitch periodicity. Thus the simple procedure of picking the largest peak in the autocorrelation function fails in these cases.

Another difficulty with the autocorrelation method is that a large amount of computation is required even for the center clipped speech. A simple modification of the center clipping function leads to great simplification of the autocorrelation function with essentially no degradation in pitch detection. This modification is shown in Figure 11. As indicated there, the output is +1 if the input is greater than C_L, and -1 if the input is less than -1. Otherwise, the output is zero. This clipping function will be called a three level center clipper.

The computation of the autocorrelation function for a three level center clipped signal is particularly simple. If we denote the output of a three level center clipper as y(n), then the product terms $y(n+m)\ y(n+m+k)$ in the autocorrelation function

$$R_n(k) = \sum_{m=0}^{N-k-1} y(n+m)\ y(n+m+k) \tag{6}$$

can have only three different values

$$\begin{aligned} y(n+m)\ \ y(n+m+k) &= 0, \text{ if } y(n+m) = 0,\ y(n+m+k) = 0 \\ &= +1, \text{ if } y(n+m) = y(n+m+k) \neq 0 \\ &= -1, \text{ if } y(n+m) \neq y(n+m+k) \end{aligned} \tag{7}$$

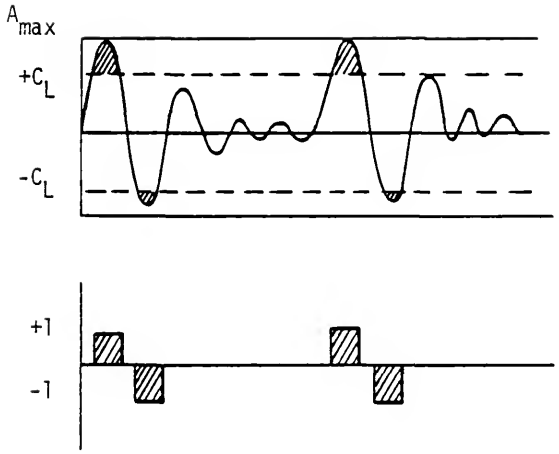


Figure 11. Illustration of the 3-level center clipper.

Thus, all that is required is some simple combinatory logic and increment/decrement instead of multiplications.

We have implemented the modified autocorrelation pitch detection algorithm in a computer program "AUTOC." A three level center clipper with clipping level set to 65% of maximum value is used. The autocorrelation is computed and the peak location decided. If the peak value exceeds a threshold value, then the speech segment is voiced; otherwise, it is unvoiced. The threshold value is usually chosen as 30% of $R_n(0)$.

Comparison of the Two Algorithms

We have tested both algorithms on real speech data. One typical result is shown in Figure 12. The speech utterance is a sentence "We were away a year ago" by a child subject. The silence interval is represented by a pitch period value of -5. The unvoiced speech segment is represented by a pitch period value of 0. Otherwise, the speech segment is voiced with the pitch period shown.

One interesting feature of this subject's speech is that there are sudden jumps in pitch either at the beginning or at the end of voicing intervals. This phenomenon is evidenced by the speech waveform in Figure 13. The top figure shows the speech waveform at the beginning of voicing. The bottom figure shows the speech waveform at the end of voicing.

Now let us compare the pitch period contours in Figure 12. We can see that the two algorithms give the same result over a large portion of voiced speech. The major difference between the two algorithms occurs at the beginning or the end of voicing intervals where the pitch period has changed suddenly. We see that the modified autocorrelation method has followed the change in pitch period, while the cepstrum method did not follow the pitch change and labeled the voiced segment as unvoiced.

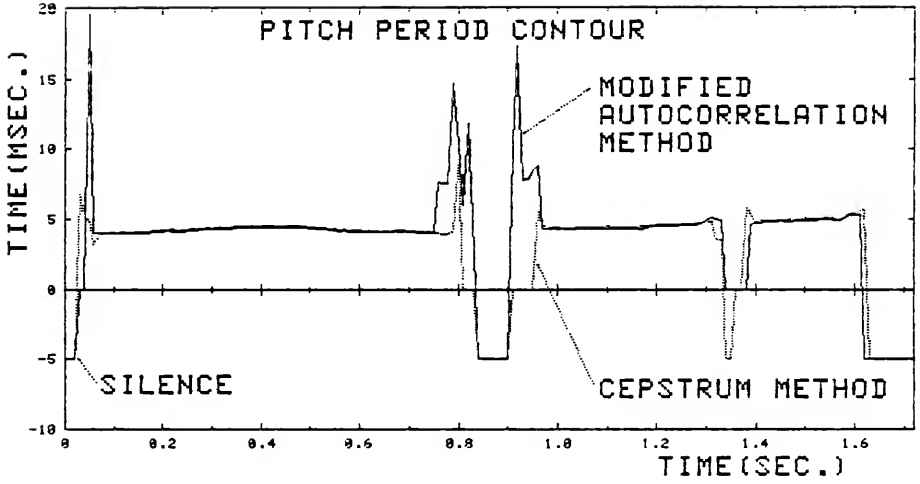


Figure 12. Comparison of the performance of the two pitch detection algorithms.

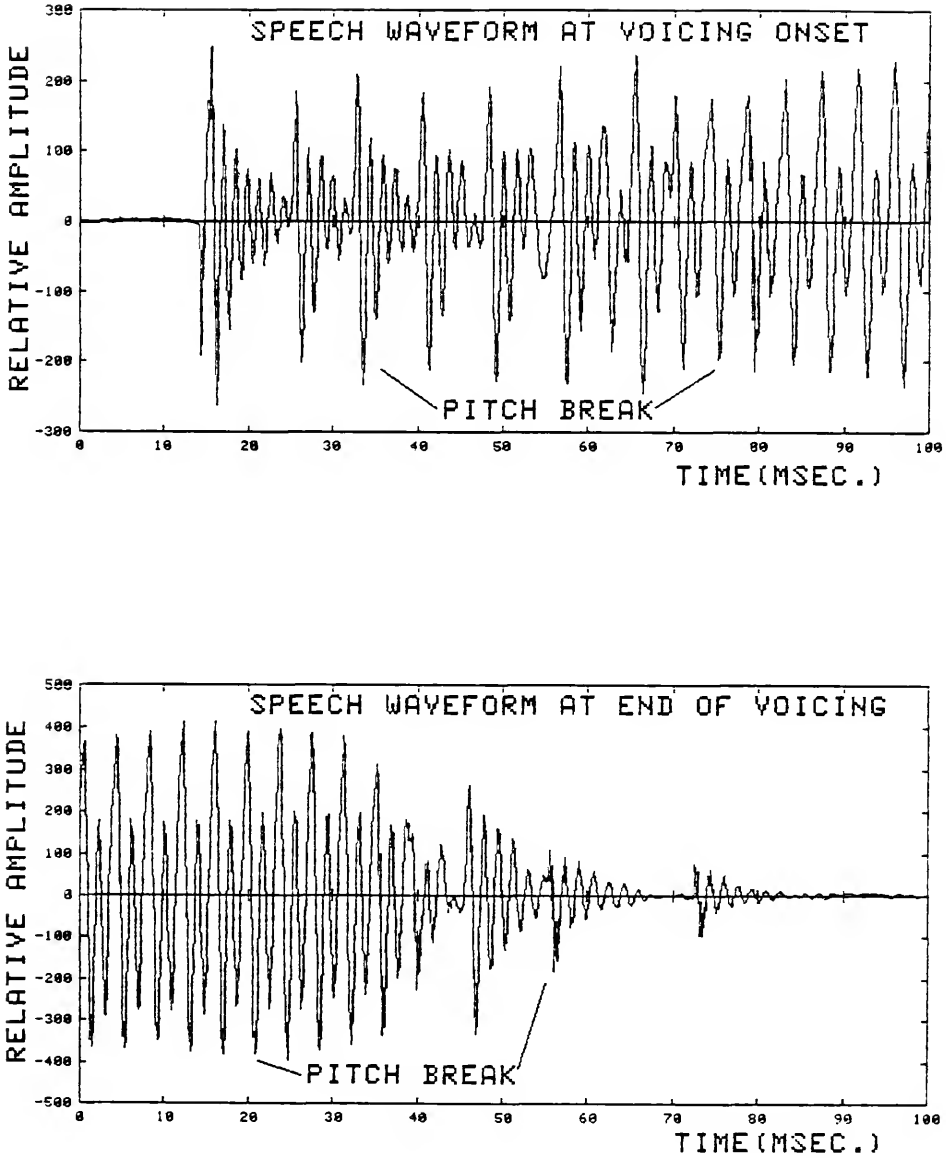


Figure 13. Speech waveform at the beginning (top) and the end of voicing showing the "pitch break" phenomena.

The problem with the cepstrum method is that it is a frequency domain approach. When there is a rapid change in pitch, the ripples in the logarithm spectrum are weakened, and so is the pitch peak in the cepstrum. Thus the pitch period is usually not correctly determined.

Because of the difficulty of the cepstrum method in the above situations, we decided to use the modified autocorrelation method in this research.

Formant Analysis

Methods

Formant data are the most important parameters in distinguishing different speech sounds. Thus formant analysis is useful not only for speech synthesis purposes but also for speech recognition. The two techniques being widely used at present for formant estimation are based on cepstral analysis [22] and linear prediction [23,24]. We have implemented the linear prediction formant analysis algorithm because it offers the advantage of minimal computation and maximal accuracy in formant estimation. The algorithm we implemented is based on MacCandless' algorithm [24].

Linear prediction analysis is a way to estimate the transfer function of the vocal tract filter. According to the theory of linear prediction, a sample of speech can be predicted from a linear combination of M previous samples of speech,

$$s(n) = \sum_{k=1}^{M} a_k \; s(n - k) + e(n) \qquad (8)$$

where e(n) is the error of prediction. Expressed in terms of the Z-transform, the prediction equation becomes

$$S(z) = \sum_{k=1}^{M} a_k \; S(z) \; z^{-k} + E(z) \qquad (9)$$

and the transfer function between S(z) and E(z) is

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{M} a_k z^{-k}} = \frac{1}{A(z)} \tag{10}$$

where the denominator polynomial $A(z) = 1 - \sum_{k=1}^{M} a_k z^{-k}$ is usually called the predictor polynomial.

Because the basic property of linear prediction requires that the error sequence, e(n), has a flat spectral envelope, the function H(z) can provide a good estimate of the transfer function of the vocal tract filter. There are two ways of estimating the formants using H(z). One is to compute the frequency response of H(z),

$$\left|H(e^{j\omega T})\right| = \frac{1}{\left|A(e^{j\omega T})\right|} = \frac{1}{\left|1 - \sum_{k=1}^{M} a_k e^{-j\omega kT}\right|} \tag{11}$$

where T is the sampling frequency of speech. The peaks in the frequency response, $|H(e^{j\omega T})|$, can then be decided and used as estimates of the formant locations. Another approach is to find the poles of H(z), or equivalently the roots of the polynomial A(z). The frequencies and bandwidths of the complex pole pairs will then correspond to the frequencies and bandwidths of the formants. To be more specific, the formant frequency (F) and bandwidth (B) corresponding to the complex pole P is given by

$$F = \frac{1}{2\pi T} \text{ Im}(\log P) \quad \text{and} \quad B = -\frac{1}{\pi T} \text{ Re}(\log P) \tag{12}$$

where Im(·) and Re(·) represent imaginary part and real part, respectively. The block diagram of formant estimation is given in Figure 14.

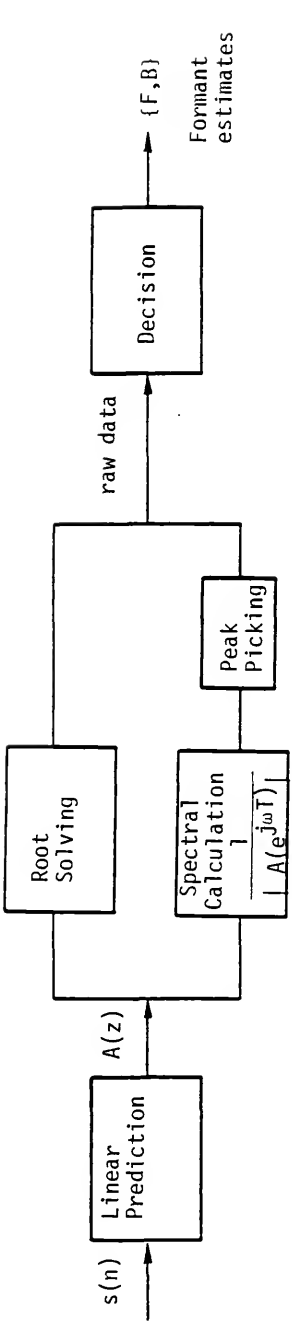


Figure 14. A general procedure for formant estimation.

Notice that in Figure 14 there is a block labeled "decision" before the formant estimates are obtained. This is because the number of spectral peaks (or complex pole pairs of H(z)) usually is not equal to the number of formants. In the case of the root extraction method, the number of complex pole pairs exceeds the number of formants because additional poles are needed to account for the source characteristics. In the peak-picking method, merging of closely-spaced peaks can also cause a problem in formant estimation. Thus the additional decision logic is needed to assign peaks (or poles) to formants.

The most common way of assigning peaks (or poles) to formants is using the "continuity" constraint. This means the formant frequencies of adjacent speech frames (usually defined as 20 ms of speech) cannot differ by a large amount because of the physical constraint of the human vocal system. Thus if the formant estimations of the neighboring speech frame are available, the formant estimates of the current segment can be obtained by assigning the peak (or pole) to the formant that is closest in frequency. This is the essence of the McCandless algorithm.

The details of the McCandless algorithm are as follows. We will discuss the algorithm for the peak-picking method only. The algorithm for the root extraction method is very similar, except that the bandwidth value will be utilized rather than peak amplitude.

Since the continuity constraint is used to assign peaks to formants, we have to make sure that the initial estimates of formant locations are correct. Otherwise, the formant trajectory may follow a wrong track. The correctness of the initial formant estimates is guaranteed by starting the processing at the stationary portion of a vowel (called anchor point), and then branching outward in both directions using the most recent formant frequency estimates as the next reference. The anchor points can be defined

automatically using the voicing and intensity information. In our implementation, however, the anchor points are defined by the users. This makes the program more flexible, and thus more accurate results will be obtained.

Once the anchor point is defined, the analysis will proceed in the manner shown in Figure 15. Processing of the backward branch is begun at the next anchor point and continued until an unvoiced frame is encountered, or until a frame is encountered which has already been processed by a previous forward branch. Then the forward branch from the same anchor point is begun and continued until an unvoiced frame is encountered, or until a new subdivision boundary is reached. At this point, processing jumps to the next anchor point and begins again with a backward branch, and so forth, until the processing is complete. Notice that the unvoiced region is not processed.

At each frame peaks in the spectral envelope, P_i's, must be mapped into formants, F_i's. This is done by the following steps.

Step 1: Fill Slots. Fill each formant slot S_i, i = 1 to 4, with the best candidate P_j by placing the peak P_j closest in frequency to the estimates EST_i into slot i.

Step 2: Remove Duplicates. If the same peak P_j fills more than one slot S_i, keep it only in the slot S_k which corresponds to the estimate EST_k closest in frequency, and remove it from any other slots.

Step 3: Deal with Unassigned Peaks. If all peaks P_j have been assigned to formant slots, go to Step 4. Otherwise try to fill empty slots with values not assigned in Step 1 as follows:

(a) If there is an unassigned peak P_k, and an unfilled slot S_k, fill the slot with the peak and go to Step 4. If P_k is unassigned, but slot S_k is filled, check the amplitude (amp) corresponding to P_k as follows:

EST_i = Frequency estimates for four formants in current frame.

P_i = Frequency locations of peaks in current frame.

F_i = Formant frequencies decided in current frame.

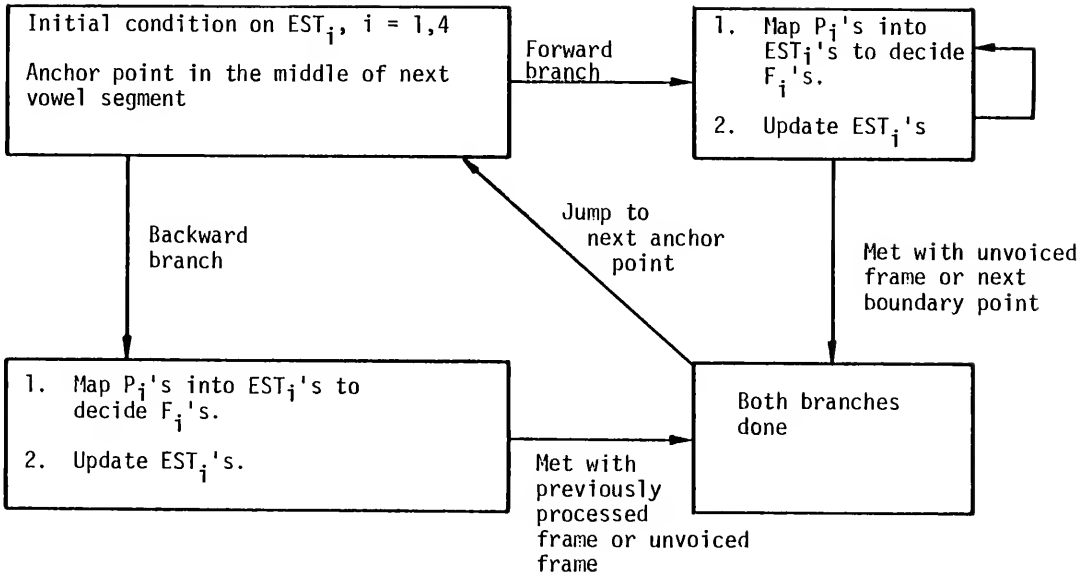


Figure 15. Flow chart of anchor point scheme.

if amp(P_k) < $\frac{1}{2}$ · amp(peak assigned to S_k) throw P_k away and go to Step 4. Otherwise, go to (b).

(b) If P_k is unassigned, and S_{k+1} is unfilled, move the peak in S_k to S_{k+1} and put P_k in S_k. Go to Step 4.

(c) If P_k is still unassigned, but S_{k-1} is unfilled, move the peak in S_k to S_{k-1}, and put P_k in S_k. Go to Step 4. If (a), (b), and (c) all fail, ignore P_k.

Step 4: Deal with Unfilled Slots. If S_1, S_2, and S_3 are all filled, go to Step 5. (F_4 may or may not be filled.) Otherwise, recompute the spectrum on a circle with radius less than unity to hopefully separate merged peaks. Fetch the peaks and go to Step 1.

Step 5: Update Estimates. Accept formant slot contents as formant estimates for this frame, i.e., $F_i = S_i$, i = 1, 2, 3. Also, use formant slot contents as estimates for next frame, i.e., $EST_i = S_i$, i = 1, 2, 3, 4. (If a slot is empty, keep the original formant estimate for that formant.)

There will still be the possibility that a formant slot has not been filled or that the formant value is grossly out of line for one or several frames. Thus, a final smoothing procedure is needed to resolve these situations. We will not go into the details of final smoothing here.

Results and Discussion

Both the root extraction method and the peak picking method have been implemented as computer programs. In the peak picking method, the formant frequency and amplitude data are estimated. The root extraction method, on the other hand, analyzes the formant frequency and bandwidth. The formant amplitude information is useful in parallel formant synthesis. The formant bandwidth information is useful in cascade formant synthesis.

The two programs were used to analyze one test sentence. The results are shown in Figure 16. The sentence is "It is a bird." We see that the two methods give similar results. Since the root-extraction method obtains the formant bandwidth information which is important for synthesizing high quality speech, we will use this method in our research.

Glottal Inverse Filtering

According to the acoustic model of speech production, the speech signal is the output of a linear time-invariant filter. The model of speech production is drawn in Figure 17 in block diagram form. Here we restrict our attention to voiced sounds, so the excitation to the vocal tract filter is the glottal volume velocity.

According to linear system theory, it is possible to reverse the process of speech generation to obtain the glottal volume velocity. This process is called glottal inverse filtering. The glottal volume velocity waveform can then be used in formant synthesis for high quality speech. Another application of the glottal volume velocity waveform is detection of laryngeal pathology because the glottal volume velocity is affected by the condition of the larynx [25].

The process of glottal inverse filtering is illustrated in Figure 18. The speech signal is first passed through a filter whose transfer function is the inverse of the vocal tract filter, then through an integrator which offsets the radiation effect to obtain the glottal volume velocity.

As we discussed before, the vocal tract filter can be modeled as a cascade of resonant circuits for vowel sounds. The nth resonant circuit has a transfer function (in terms of Z-transform)

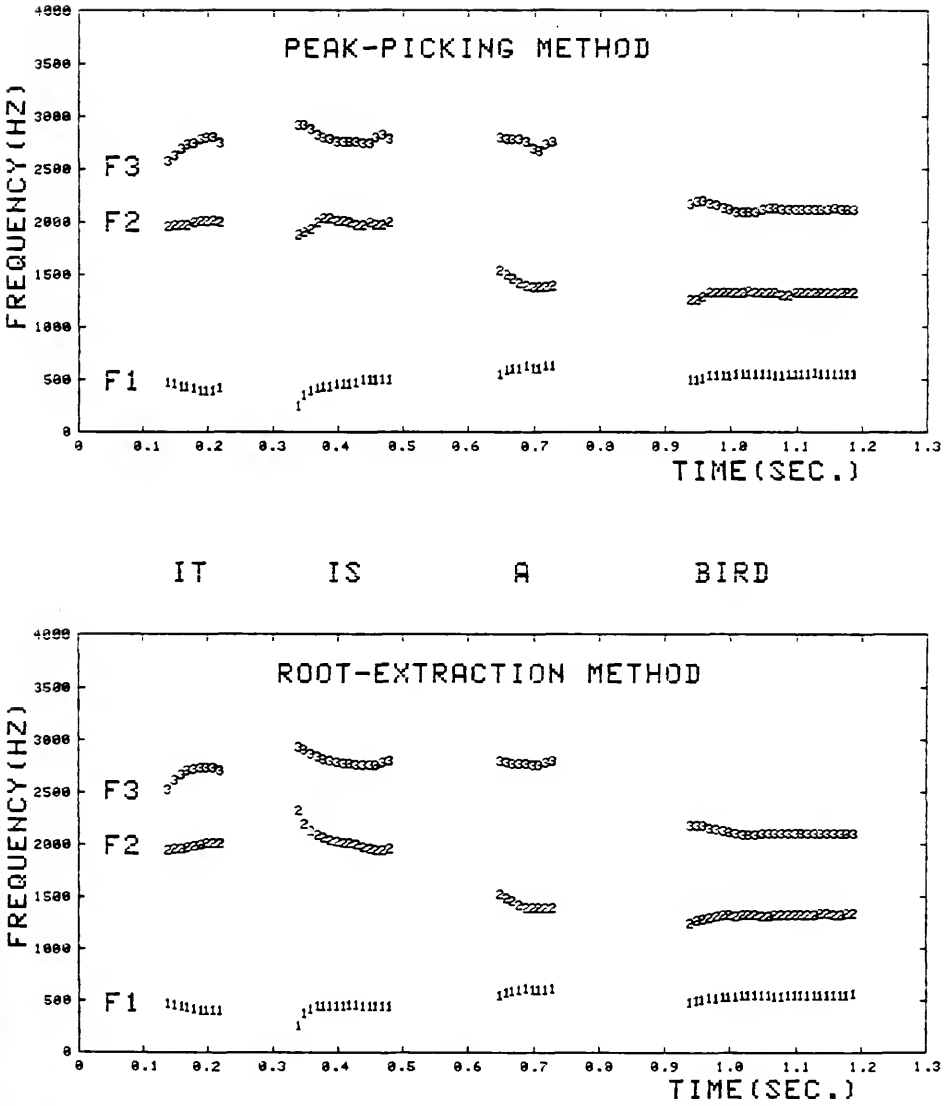


Figure 16. Comparison of results of two formant tracking algorithms.

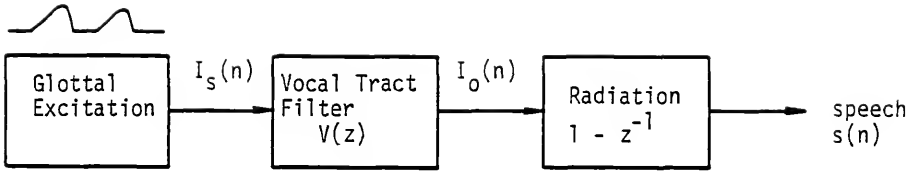


Figure 17. Model of speech production.

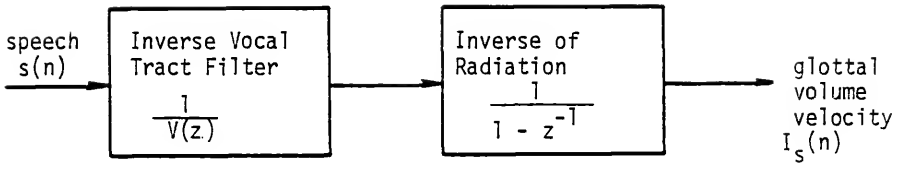


Figure 18. Block diagram of glottal inverse filtering.

$$V_n(z) = \frac{1 - 2e^{-\pi B_n T} \cos 2\pi F_n T + e^{-2\pi B_n T}}{1 - 2e^{-\pi B_n T} \cos 2\pi F_n T\, z^{-1} + e^{-2\pi B_n T}\, z^{-2}} \qquad (13)$$

where F_n is the nth formant frequency and B_n is the nth formant bandwidth. Notice that the numerator constant is chosen such that the DC gain is 0 dB. The overall transfer function for the vocal tract filter is the product of these terms

$$V(z) = K \prod_{n=1}^{M} V_n(z) = K \prod_{n=1}^{M} \frac{1 - 2e^{-\pi B_n T} \cos 2\pi F_n T + e^{-2\pi B_n T}}{1 - 2e^{-\pi B_n T} \cos 2\pi F_n T\, z^{-1} + e^{-2\pi B_n T}\, z^{-2}} \qquad (14)$$

where M is the number of formants needed to account for the transmission characteristics of the vocal tract filter in the frequency range of interest, and K is a gain constant. The transfer function V(z) has poles only, so the vocal tract filter for the vowels is an all-pole filter. The transfer function of the inverse vocal tract filter is given by

$$A(z) = \frac{1}{V(z)} = \frac{1}{K} \prod_{n=1}^{M} \frac{1 - 2e^{-\pi B_n T} \cos 2\pi F_n T\, z^{-1} + e^{-2\pi B_n T}\, z^{-2}}{1 - 2e^{-\pi B_n T} \cos 2\pi F_n T + e^{-2\pi B_n T}} \qquad (15)$$

Thus, the inverse vocal tract filter is an all zero filter which is guaranteed to be stable.

In order to construct the inverse vocal tract filter, we have to know the formant frequencies and bandwidths. For the glottal inverse filtering problem, the estimation of the formant frequencies and bandwidths should be handled with great caution. The integration operation in the glottal inverse filtering process will amplify small errors in the inverse vocal tract filter and result in a distorted glottal volume velocity waveform. An accurate method of estimating the formant frequencies and bandwidths is called the closed-phased analysis [26]. This method applies the LP analysis

in an interval of speech where the glottis is closed (or the glottal volume velocity is zero). Since there is no excitation during this interval, the linear prediction model is exact and so is the analysis result. The definition of the closed-phased period and its relationship to the speech waveform is illustrated in Figure 19. Notice that the largest negative peak in the speech waveform usually occurs at the instant the glottal volume velocity becomes zero. Thus, closed-phase analysis can be applied to the speech signal after this instant. The roots of the predictor polynomial obtained by the closed-phase analysis are then used as an estimate of formant frequencies and bandwidths.

The inverse filtering procedure discussed above is applied to a real speech signal. Figure 20 shows a typical result. The vowel is /OW/ as in "rose" spoken by a female subject. The LP analysis is applied to the speech segment from 140 samples to 170 samples which is right after the glottal closing point. The roots of the predictor polynomial are shown in the middle of the figure. Except for the roots with frequency equal to 0 and 3193.10 Hz, all the other roots correspond to formants. We used these formants to construct the inverse filter. The bottom figure shows the results of inverse filtering. The solid line is the glottal volume velocity waveform and the dotted line is the output (residue) of the inverse vocal tract filter.

There are two things worth noting in the glottal waveform. They are

(a) The flat portion of the glottal volume velocity is not located at the absolute zero. This is because the differentiation effect introduced by radiation has destroyed the zero reference for the volume velocity waveform.

(b) The flat portion of the glottal volume velocity is about 10 samples (1 ms) long, while we have used 30 samples in the closed-phase analysis.

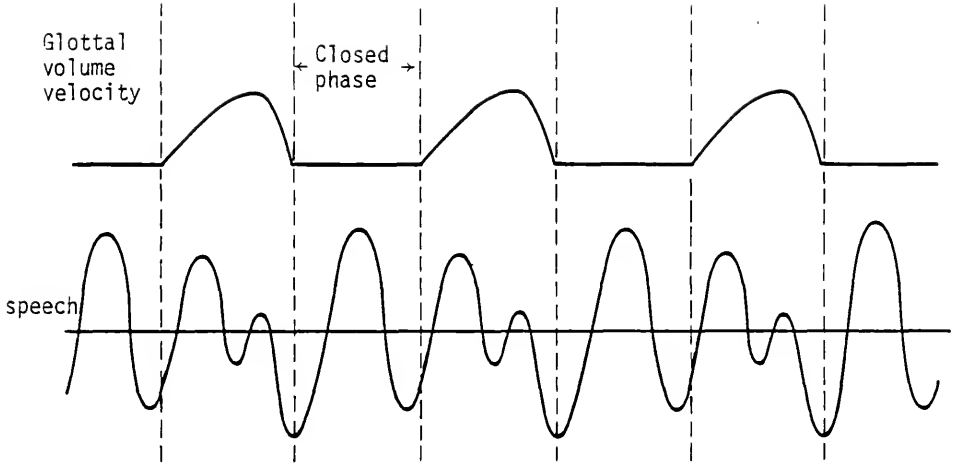
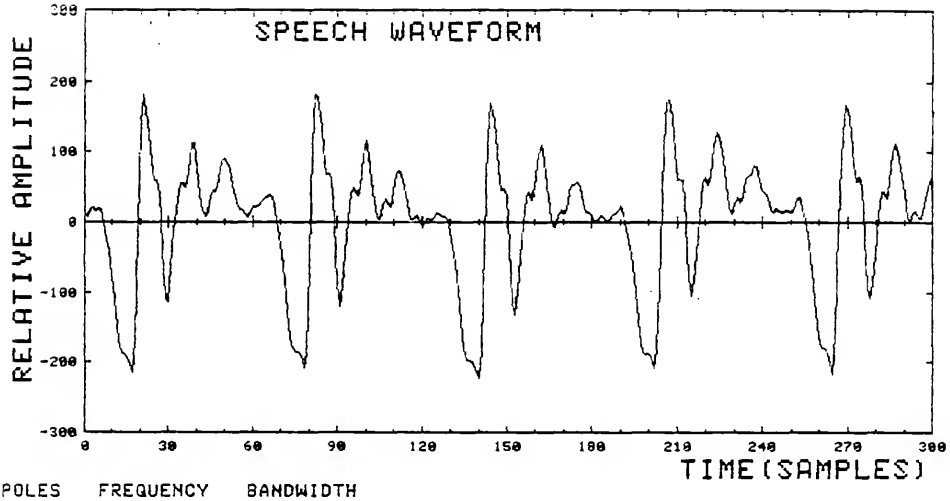
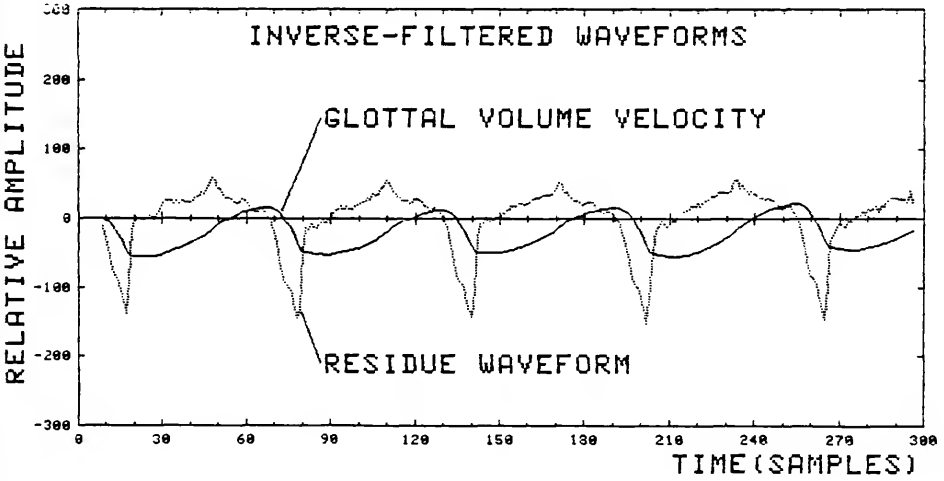


Figure 19. The definition of the closed-phase interval and its relationship to the speech waveform.

Figure 20. An example of inverse filtering. The speech waveform is shown in the top figure. The pole locations are shown in the middle. The bottom figure shows the inverse-filtered waveforms.



POLES	FREQUENCY	BANDWIDTH
1	0.00	6593.98
2	0.00	2030.36
3	-3193.10	2673.08
4	3193.10	2673.08
5	-669.85	191.58
6	669.85	191.58
7	-3747.16	440.46
8	3747.16	440.46
9	-1608.64	176.04
10	1608.64	176.04
11	-2697.32	127.10
12	2697.32	127.10



Thus the assumption of closed-phase analysis is not satisfied. This situation is usually found in female or child speech. In some cases, the glottal volume velocity waveform will be distorted; in other cases (as in the example) a good volume velocity waveform can still be obtained.

As discussed before, the glottal inverse filtering procedure is very sensitive to low frequency distortion because of the integration operation. One source of distortion is the phase distortion introduced by the tape recorder. Another type of distortion is the 60 Hz noise due to imperfect grounding. Our solution to both problems is to filter the frequency components below 60 Hz completely.

The glottal volume velocity waveform can be used as the excitation function for a formant synthesizer. Also, other types of excitation functions can be derived from the glottal volume velocity function. We will discuss the details of constructing the glottal excitation functions in Chapter 3.

Analysis of a Sentence for Synthesis

Since the main objective of this research is speech synthesis, we will conclude this chapter by discussing an example of how we analyze a sentence to extract the necessary parameters for synthesis. The speech analysis procedure is shown in the block diagram in Figure 21. The necessary parameters for speech synthesis are the intensity, the pitch/voicing information, the formant frequency/bandwidth, and the glottal volume velocity. Except for the intensity, the analysis of the other parameters has been discussed in the previous sections. The intensity is, by definition, the root-mean-square (rms) value of the speech. Thus, the intensity contour for a sentence can be obtained by computing the rms value of every

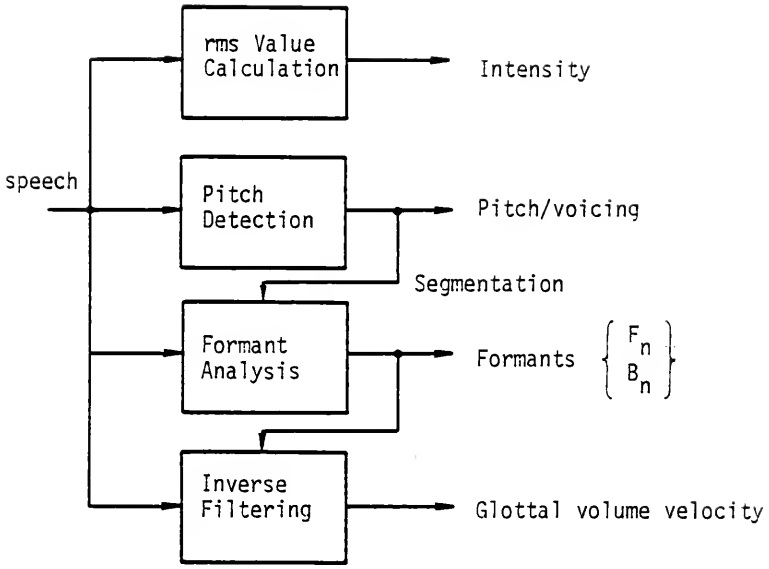


Figure 21. Summary of the speech analysis procedure.

10 ms segment of speech. To smooth out the discontinuity in the intensity contour due to the positioning of the speech segment, we have applied a 30 ms Hanning window* on the speech segment. The Hanning window is centered around the speech segment as illustrated in Figure 22.

The analysis procedure is applied to a sentence "It is a bird." This sentence was used in our previous studies [27,28]. The dialog of using the analysis program is shown in Appendix A. The analysis results are shown in Figure 23. They are, from top to bottom, the intensity contour, the pitch-period contour, and the formant frequency contours. In comparing the intensity contour with the pitch period contour, we can find a match between the low intensity portion and the unvoiced/silence region. This indicates that the fricatives are usually of lower intensity than vowels. Another feature of the pitch period contour is that there is a break in pitch at the end of the sentence. This phenomenon is also observed in the spectrogram in Figure 24 where the spacing between the vertical striation is proportional to the pitch period.

The formant frequency contours are shown in the bottom of Figure 23. The formant frequency contours are similar to the spectrogram in Figure 24. The formant bandwidths are also estimated. We did not show them here because the formant bandwidth estimation is usally less accurate than the formant frequency estimation. Also, the formant bandwidth is not guaranteed to be continuous, so the bandwidth contours may jump up and down and cross each other. This makes it difficult to interpret the formant bandwidth information. In our synthesis, we usually have to correct the formant bandwidth values by applying the closed-phase analysis.

---

*Hanning window is defined by $W(n) = \begin{cases} 1/2(1 - \cos(2\pi n/L)) & 0 \le n \le L \\ 0 & \text{other} \end{cases}$

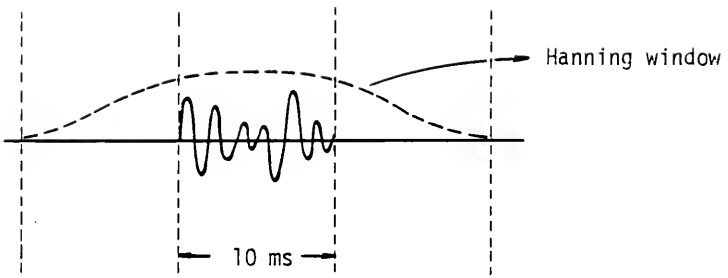


Figure 22. Application of Hanning window for estimating the intensity contour.

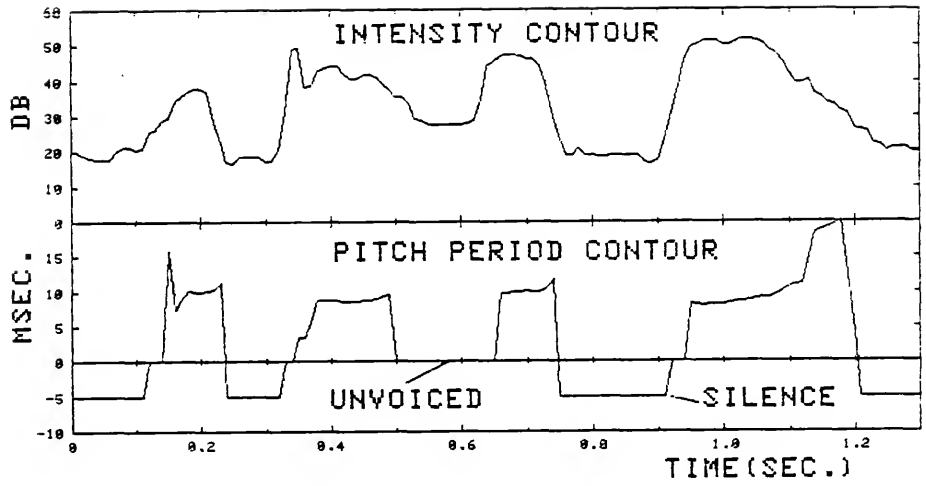


Figure 23. Intensity, pitch (top), and formant contours (bottom) of the sentence "It is a bird."

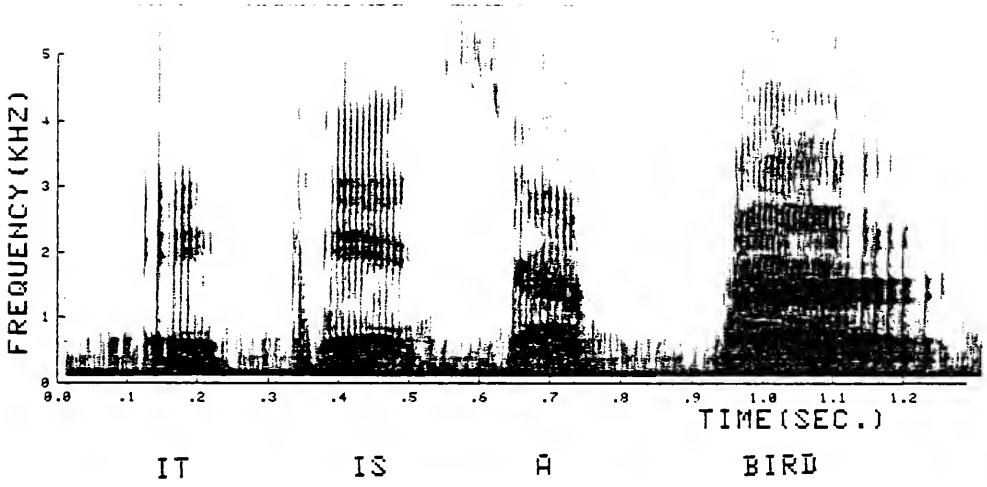


Figure 24. Spectrogram of the sentence "It is a bird."

CHAPTER 3
SPEECH SYNTHESIS

This chapter concerns the synthesis of natural sounding speech with different types of glottal excitation functions. In the first section we discuss the speech synthesizer configuration, i.e., a serial/parallel formant synthesizer. We will illustrate the basic principles of speech synthesis by synthesizing simple vowels and consonants. The next section is devoted to the discussion of different glottal excitation models and the derivation of these models from speech. The third section describes the speech data used in this research, i.e., three different sentences produced by a male, a female, and a child, respectively. These sentences were analyzed to extract parameters necessary for our synthesis study. The sentences were then resynthesized using different glottal excitation functions. We will discuss the difficulties we encountered in synthesizing each sentence. We also compare the differences between male, female, and child speech. Finally, the spectrograms of the natural and synthetic speech are compared.

A Cascade/Parallel Formant Synthesizer

The formant synthesizer used in our research is based on the cascade/parallel formant synthesizer by Klatt [16]. The Klatt synthesizer was first implemented on the Nova 4 computer by B. George for her master's thesis [29]. Later we modified the glottal excitation portion of the original synthesizer which will be discussed in detail in the next section. For the present,

we are going to discuss the overall configuration of the synthesizer and the basic principles of synthesizing speech.

The block diagram of the cascade/parallel formant synthesizer is shown in Figure 25. The two major components of the speech synthesizer are the source excitation and the vocal tract filter. There are three source excitations: the voicing source for voiced sounds such as vowels, the aspiration source for synthesizing the aspirative, and the frication source for synthesizing the fricatives such as /s/, /f/, etc. The aspiration source and frication source are simply random noise generators. The voicing source is more complex and will be discussed in the next section.

The vocal tract filter consists of two portions. The first is a cascade resonator branch which represents the vocal tract transfer function for laryngeal sources (voicing source and aspiration source). It has been shown by Fant [30] that for the cascade connection, the relative amplitudes of formant peaks for vowels will come out just right without the need for individual amplitude controls. The details of the cascade branch are shown in Figure 26, where R1 to R5 represent the five formant resonators. Notice that there is an additional RNZ-RNP pair which represents a resonance/anti-resonance pair for synthesizing nasalized sounds. As for the non-nasal sounds, the frequency of RNZ is set equal to the frequency of RNP, so the two cancel each other.

The second portion is a parallel resonator branch which represents the vocal tract transfer function for the frication source. The detailed configuration is shown in Figure 27. Notice that each resonator has an amplitude control which adjusts the relative amplitude of the formant peak. As illustrated in Equation 16, the summation of the two all-pole functions will create additional zeros, so the summation of the resonator transfer

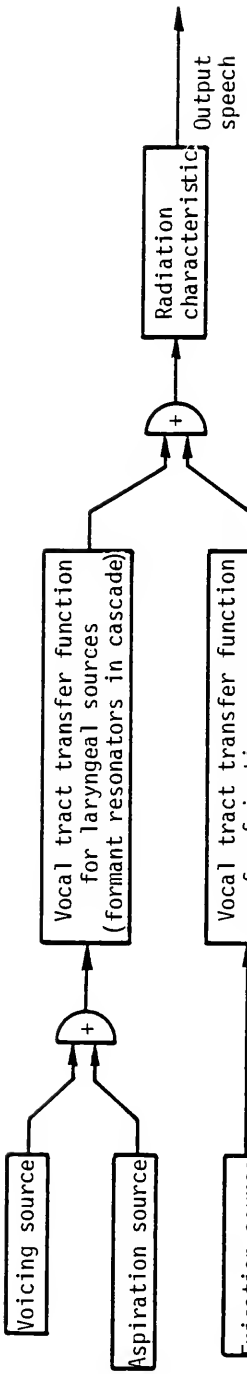


Figure 25. The cascade/parallel configuration of the formant synthesizer.

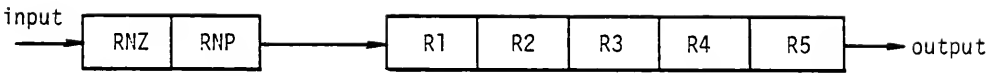


Figure 26. Configuration of cascade branch.

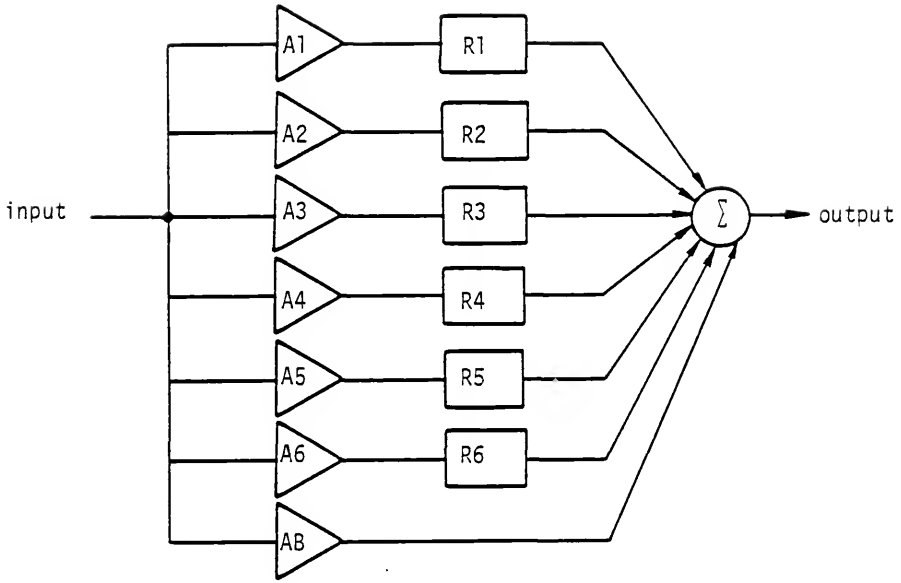


Figure 27. Configuration of the parallel branch.

$$\frac{1}{z - p_1} + \frac{1}{z - p_2} = \frac{2(z - \frac{p_1 + p_2}{2})}{(z - p_1)(z - p_2)} \qquad (16)$$

functions will create antiresonances. Thus with proper amplitude controls, the parallel branch is capable of approximating any rational transfer function. This makes the parallel branch suitable for synthesizing fricatives which have both poles and zeros in their transfer functions. Other features of the parallel branch are the addition of a 6th formant resonator (R6) and a bypass path with amplitude control AB. The 6th formant frequency is fixed at 4900 Hz for the synthesis of very high frequency noise in [s, z]. The bypass path is present because the transfer functions for [f,v,θ,ζ,p,b] contain no prominent peaks, and the synthesizer should include a means of bypassing all the resonators to produce a flat transfer function.

Besides the sound sources and the vocal tract filter, a first order difference filter is needed to simulate the effect of the radiation. This completes our discussion of the cascade/formant synthesizer. The detailed configuration of the synthesizer is shown in Figure 28, and the parameters are summarized in Table 1.

The cascade/parallel formant synthesizer has been implemented as an interactive speech synthesis program "HANDSY3" on a Nova 4 minicomputer. The program is capable of doing the following functions:

1) Reading parameter values from an analysis file,
2) Modifying parameter values interactively on a graphics terminal, and
3) Having parameter values entered from the keyboard.

Thus the program can be used in both formant vocoder studies and in synthesis-by-rule studies. The dialog of using the interactive speech synthesis program is listed in Appendix A with a graphic plot illustrating the parameter modification process.

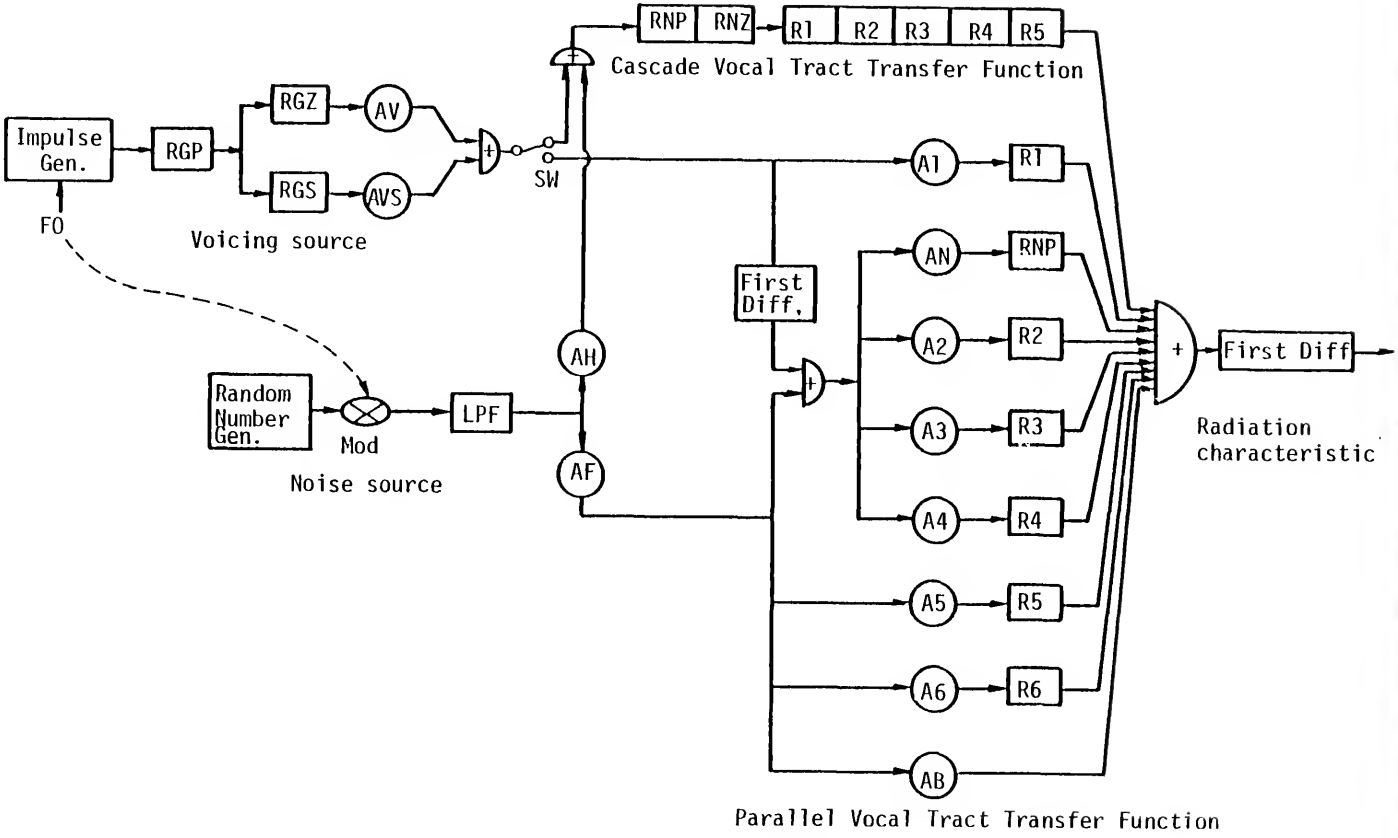


Figure 28. Detailed block diagram of Klatt's formant synthesizer.

Table 1. List of control parameters for Klatt's formant synthesizer.

N	V/C	Sym	Name	Min	Max	Typ
1	V	AV	Amplitude of voicing (dB)	0	80	0
2	V	AF	Amplitude of frication (dB)	0	80	0
3	V	AH	Amplitude of aspiration (dB)	0	80	0
4	V	AVS	Amplitude of sinusoidal voicing (dB)	0	80	0
5	V	F0	Fundamental freq. of voicing (Hz)	0	500	0
6	V	F1	First formant frequency (Hz)	150	900	450
7	V	F2	Second formant frequency (Hz)	500	2500	1450
8	V	F3	Third formant frequency (Hz)	1300	3500	2450
9	V	F4	Fourth formant frequency (Hz)	2500	4500	3300
10	V	FNZ	Nasal zero frequency (Hz)	200	700	250
11	C	AN	Nasal formant amplitude (dB)	0	80	0
12	C	A1	First formant amplitude (dB)	0	80	0
13	V	A2	Second formant amplitude (dB)	0	80	0
14	V	A3	Third formant amplitude (dB)	0	80	0
15	V	A4	Fourth formant amplitude (dB)	0	80	0
16	V	A5	Fifth formant amplitude (dB)	0	80	0
17	V	A6	Sixth formant amplitude (dB)	0	80	0
18	V	AB	Bypass path amplitude (dB)	0	80	0
19	V	B1	First formant bandwidth (Hz)	40	500	50
20	V	B2	Second formant bandwidth (Hz)	40	500	70
21	V	B3	Third formant bandwidth (Hz)	40	500	110
22	C	SW	Cascade/parallel switch	0(CASC)	1(PARA)	0
23	C	FGP	Glottal resonator 1 frequency (Hz)	0	600	0
24	C	BGP	Glottal resonator 2 frequency (Hz)	100	2000	100
25	C	FGZ	Glottal zero frequency (Hz)	0	5000	1500
26	C	BGZ	Glottal zero bandwidth (Hz)	100	9000	6000
27	C	B4	Fourth formant bandwidth (Hz)	100	500	250
28	V	F5	Fifth formant frequency (Hz)	3500	4900	3750
29	C	B5	Fifth formant bandwidth (Hz)	150	700	200
30	C	F6	Sixth formant frequency (Hz)	4000	4999	4900
31	C	B6	Sixth formant bandwidth (Hz)	200	2000	1000
32	C	FNP	Nasal pole frequency (Hz)	200	500	250
33	C	BNP	Nasal pole bandwidth (Hz)	50	500	100
34	C	BNZ	Nasal zero bandwidth (Hz)	50	500	100
35	C	BGS	Glottal resonator 2 bandwidth	100	1000	200
36	C	SR	Sampling rate	5000	20000	10000
37	C	NWS	Number of waveform samples per chunk	1	200	50
38	C	GO	Overall gain control (dB)	0	80	47
39	C	NFC	Number of cascaded formants	4	6	5

Synthesis Strategy

Now let us discuss the principles of synthesizing speech using the cascade/parallel formant synthesizer. We will illustrate these principles by synthesizing simple vowels and consonants. The synthesis of connected speech (words, sentences) will be considered later.

Synthesis of Vowels

The parameters that are usually varied to generate an isolated vowel are the amplitude of voicing (AV), the fundamental frequency of vocal fold vibrations (FØ), the lowest three formant frequencies (F1, F2, and F3) and bandwidths (B1, B2, and B3). The fourth and fifth formant frequencies may be varied to simulate the spectral details, but this is not essential for good intelligibility.

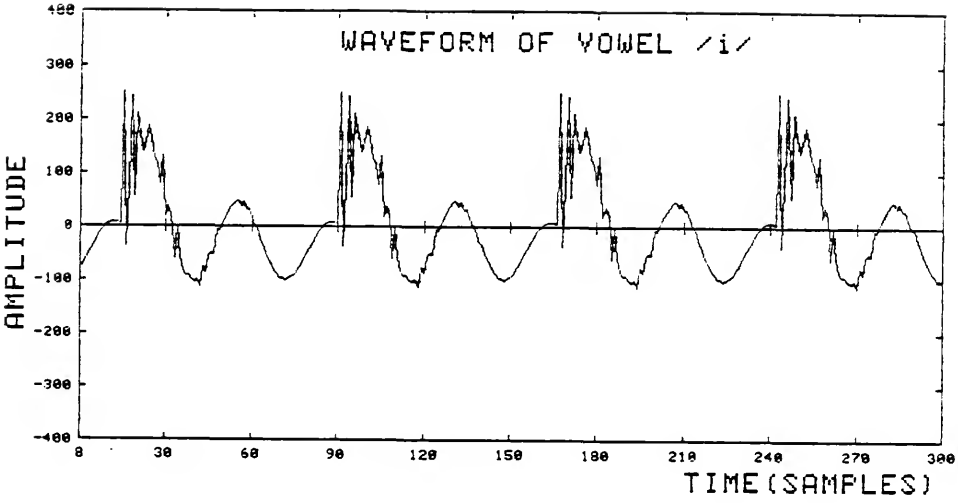
The control parameters for synthesizing the vowel /i/ are shown in Table 2. The length of the utterance is 300 ms. The amplitude of the voicing source (AV) is set to about 60 db for a stressed vowel, and falls gradually by 5 db near the end of the syllable. The fundamental frequency (FØ) follows a linear contour falling from 130 to 100 Hz. The first three formant frequencies are fixed at 310 Hz, 2020 Hz, and 2960 Hz, respectively. The first three formant bandwidths are also fixed at 45 Hz, 200 Hz, and 400 Hz, respectively. Figure 29 shows the waveform and the spectrum of the synthetic vowel.

Synthesis of Consonants

English consonants can be further categorized into fricatives, plosives, and nasals. The nasal consonant is voiced, while the fricatives and plosives can be either voiced or voiceless.

TABLE 2. Parameter Values for Vowel /i/

F1	310 Hz	B1	45 Hz
F2	2020 Hz	B2	200 Hz
F3	2960 Hz	B3	400 Hz
AV	60 dB		
FØ	130 Hz		



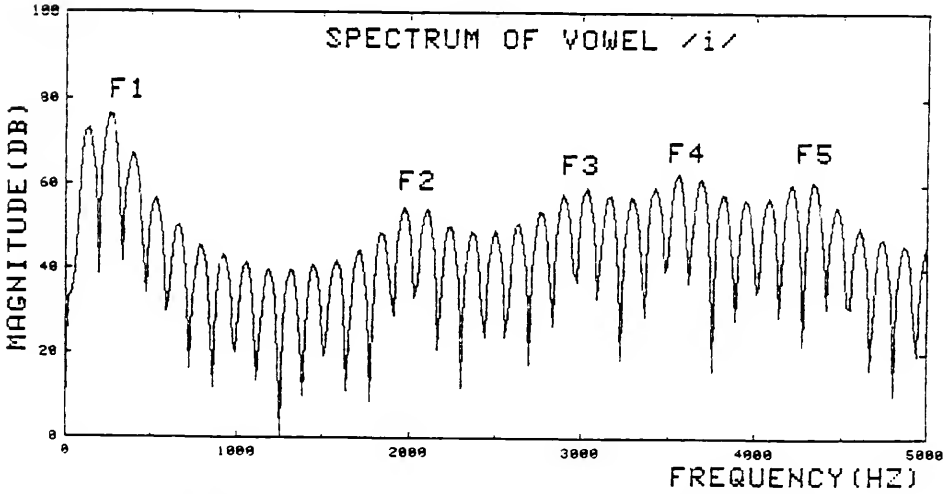


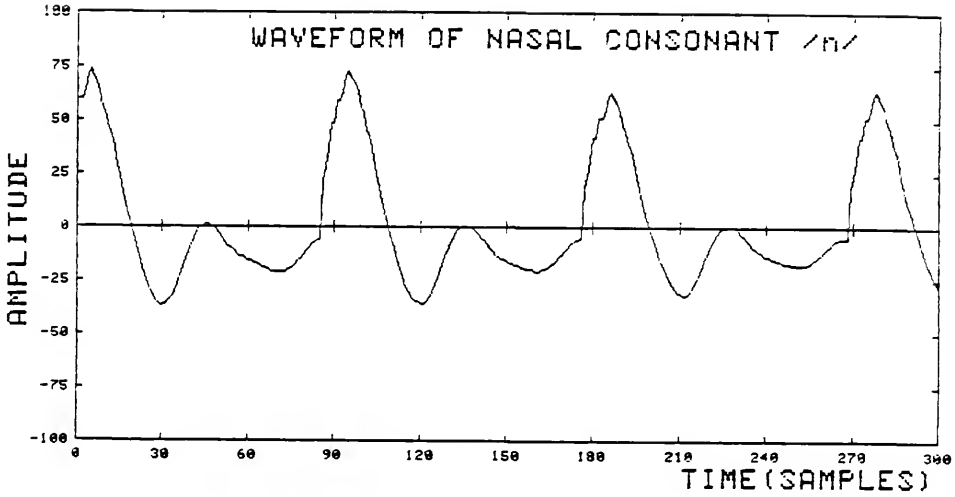
Figure 29. Waveform (top) and spectrum of the synthetic vowel /i/.

The nasal consonants are synthesized using the cascade branch. Besides the parameters used in synthesizing vowels, an additional pole-zero pair (FNP-FNZ) is added to the cascade branch for synthesizing nasal consonants or nasalized vowels. The pole-zero pair is called the nasal formant and antiformant. The nasal formant frequency is usually fixed at 270 Hz. The antiformant frequency is varied according to the degree of nasalization. For nasal consonants, the antiformant frequency is very close to the first formant frequency, so that the first formant is almost cancelled. For nasalized vowels, the antiformant frequency is equal to the mean value of the nasal formant frequency and the first formant frequency. The control parameters for synthesizing a nasal consonant /n/ are shown in Table 3. The amplitude of voicing, AV, is set at 55 dB. The fundamental frequency, FØ, is set to 100 Hz. The first three formant frequencies are 480 Hz, 1340 Hz, and 2470 Hz. The first three formant bandwidths are 40 Hz, 300 Hz, and 300 Hz, respectively. The nasal formant frequency, FNP, is 270 Hz. The nasal antiformant frequency is 450 Hz. Figure 30 shows the waveform and spectrum of the nasal consonant /n/. Notice that the nasal consonant is characterized by a dominant nasal formant peak in the spectrum.

The voiceless fricatives are synthesized using the parallel branch because the vocal tract transfer function contains both poles and zeros. The frication source, AF, is used to synthesize fricatives. The formants excited by the frication source are determined by the amplitude controls A2, A3, A4, A5, A6, and AB. The voiced fricatives are synthesized using both the voicing source, AV, and the frication source, AF. The control parameters for synthesizing a voiceless fricative /s/ are shown in Table 4.

TABLE 3. Parameter Values for Nasal Consonant /n/

F1	480 Hz	B1	40 Hz
F2	1340 Hz	B2	300 Hz
F3	2470 Hz	B3	300 Hz
FNZ	450 Hz		
FNP	250 Hz		
AV	55 dB		
FØ	100 Hz		



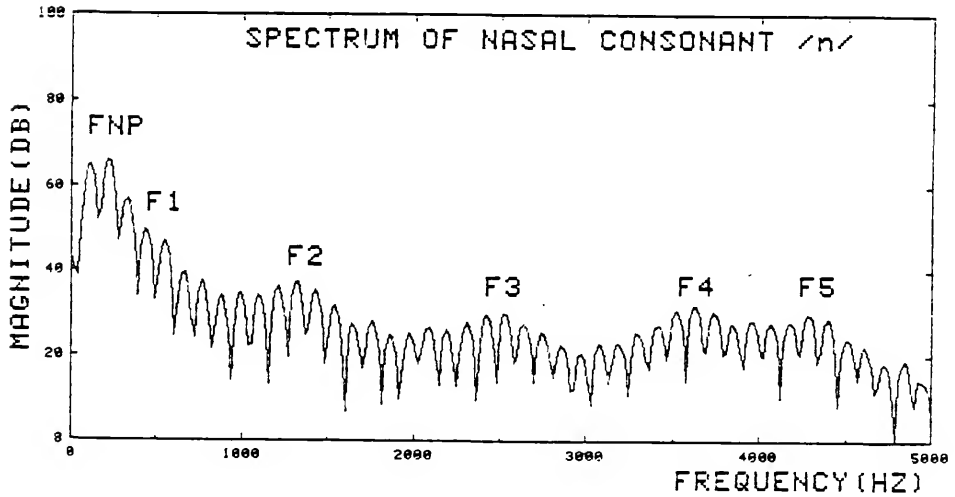


Figure 30. Waveform (top) and spectrum of the synthetic nasal consonant /n/.

TABLE 4. Parameter Values for Fricative Consonant /s/

F1	320 Hz	B1	200 Hz	A2	0 dB
F2	1390 Hz	B2	80 Hz	A3	0 dB
F3	2530 Hz	B3	200 Hz	A4	0 dB
F4	3300 Hz			A5	0 dB
F5	3750 Hz			A6	52 dB
F6	4900 Hz				
AF	40 dB				

amplitude of frication is 40 dB. For the amplitude control, A6 is set to be 52 dB; the rest are set equal to 0 dB because the fricative /s/ mainly consists of high frequency noise. The waveform and spectrum of the synthetic speech are shown in Figure 31. Notice that there are no pitch harmonics in the spectrum due to the noise-like property of the fricatives.

If we combine the fricative /s/, the vowel /i/, and the nasal /n/ with proper formant transitions, the resultant speech is the syllable /sin/. The spectrogram of the syllable /sin/ is shown in Figure 32.

Another class of consonant sounds is the plosives. The plosives are characterized by a strong sudden burst of noise and very fast formant transitions. Klatt used a step function to simulate the sudden burst of noise. We found that this will introduce a dc bias in the speech which may be perceived as low frequency noise. So, instead of using a step function, we simply change the frication amplitude, AF, from 0 to 50 db in a short time interval (say 5 ms). This will create a sudden burst of noise without any dc bias problem. The identity of the plosives is determined primarily by the formant transition patterns. Thus, to synthesize a plosive consonant we must know the formant frequencies of both the plosive and the following vowel.

We have discussed the synthesis of only a few types of consonants. There are other types of consonants such as aspiratives, affricates, etc. The basic principles of synthesizing these consonants are similar to what we have discussed above. So we will not pursue this matter further.

The Glottal Excitations for the Formant Synthesizer

The objective of this research is to study the influence of glottal excitation functions on the quality of speech. So, it is worthwhile spending some time to discuss the glottal excitation functions. Three types of glottal

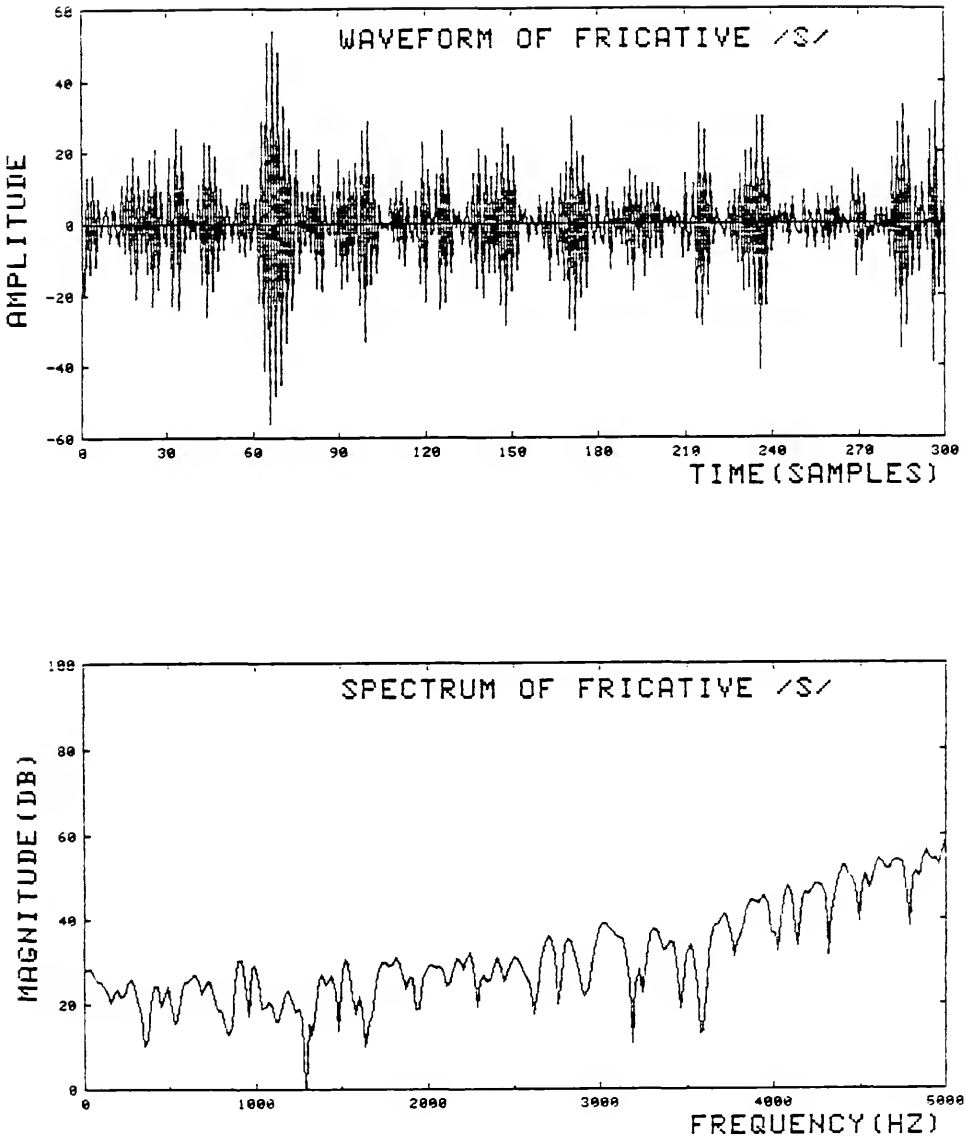


Figure 31. Waveform (top) and spectrum (bottom) of synthetic fricative /s/.

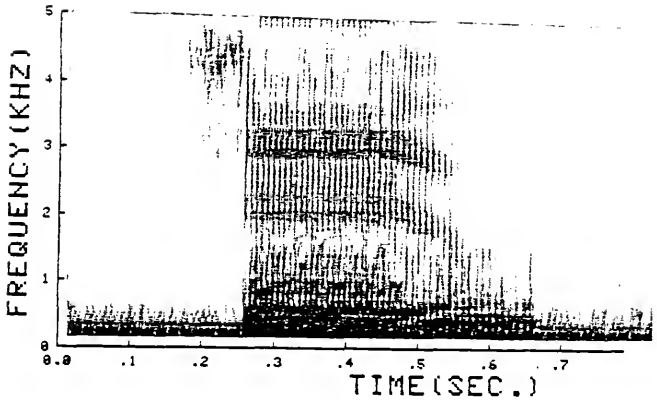


Figure 32. Spectrogram of synthesized utterance /sin/.

excitation functions are used in this research. They are

1) Impulse excitation

2) Glottal volume velocity (Fant's model)

3) Glottal area function (Guerin's model).

Impulse Excitation Source

What we call the "impulse excitation" is actually the glottal excitation of the original Klatt's synthesizer as shown in Figure 28. The impulse is first shaped by a glottal shaping filter before exciting the vocal tract filter. The block diagram of the glottal shaping filter is shown in Figure 33. The impulse train is filtered by RGP, a low pass filter. The low pass filter has a double pole with a bandwidth equal to 100 Hz. This gives the excitation function a spectrum that falls off smoothly at approximately -12 dB per octave above 50 Hz. The waveform thus generated does not have the same phase spectrum as a typical glottal pulse, nor does it contain spectral zeros of the kind that often appear in natural voicing.

The antiresonator RNZ is used to modify the detail shape of the spectrum of the excitation function with greater precision than would be possible using only a single low pass filter. It is clear from the above discussion that the impulse excitation source is a model of the spectral characteristics of the actual glottal source.

The Glottal Volume Velocity Source

The second glottal excitation function is the glottal volume velocity. As discussed in Chapter 2, the glottal volume velocity function can be derived from the speech signal by the inverse filtering process. This is the ideal glottal excitation function according to the model of speech production. But it is impractical to use this time varying glottal volume

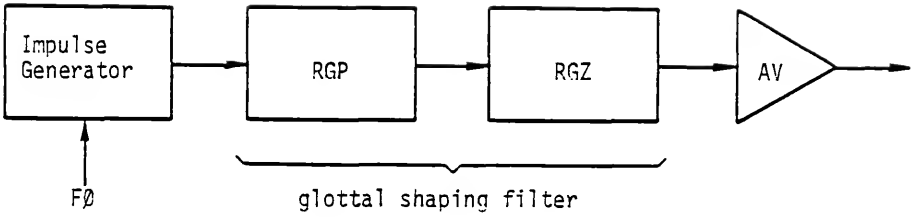
velocity for the following reasons. First, the glottal inverse filtering is very sensitive to the low frequency distortion which often occurs in the recording process. Second, it is too costly to extract the glottal volume velocity pitch synchronously because of the root extraction and the polynomial multiplication needed in the inverse filtering process.

Thus we have to use a single time-invariant glottal volume velocity waveform. Again, we cannot use the glottal volume velocity waveform derived by the inverse filtering process to synthesize sentences because the glottal volume velocity derived by inverse filtering contains formant ripples due to the effect of source-tract interaction. So the glottal volume velocity waveform differs from one vowel to another. If we use the glottal volume velocity derived from a vowel to synthesize another vowel, the synthetic speech may have an undesirable quality [9].

So the glottal volume velocity waveform used to synthesize a sentence must not contain any formant ripples (our source-tract interaction). One simple waveform which approximates the general characteristics of the glottal volume velocity and yet does not have formant ripples is given by

$$U(n) = \begin{cases} \frac{A}{2}\left(1 - \cos\frac{\pi n}{T_1}\right) & 0 \le n \le T_1 \\ A\cos\left(\frac{\pi}{2}\frac{n - T_1}{T_2}\right) & T_1 \le n \le T_2 \\ 0 & T_2 \le n \le T_0 \end{cases}$$

The waveform is shown in Figure 34, where A is the maximum volume velocity, T_0 is the pitch period, T_1 is the duration of the opening phase, and T_2 is the duration of the closing phase. Rosenberg [7] first used this waveform as glottal excitation for a formant synthesizer and showed that the resultant synthetic speech had good quality. Fant [13] also used this waveform as



$$H(z) = \frac{1}{(1 - e^{-0,01}\, z^{-1})^2} = \frac{z^2}{(z - e^{-0,01})^2}$$

Figure 33. Block diagram of the impulse excitation source.

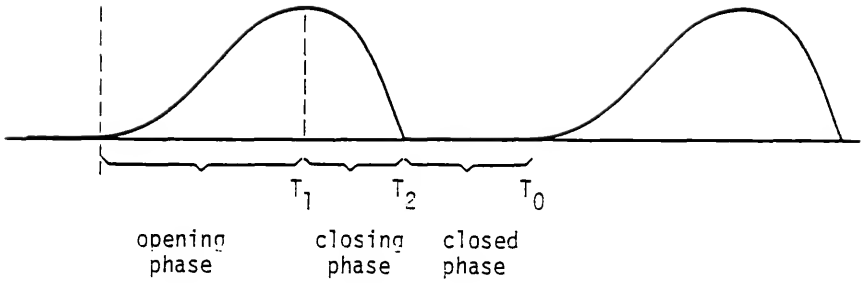


Figure 34. An idealized glottal volume velocity waveform.

a model of the glottal volume velocity and showed that it resembled the general waveshape of the actual glottal volume velocity.

Now all we have to do is to decide the waveform parameters A, T_1, and T_2. These parameters can be measured from the actual glottal volume velocity waveform.

Since the model waveform also has a -12 dB/oct fall in spectral envelope, it provides a good approximation in both time and spectral characteristics of the actual glottal volume velocity.

The Glottal Area Function Excitation Source

Rosenberg's glottal waveform model provides a good approximation of the actual glottal volume velocity except for the source-tract interaction (formant ripples). How can we incorporate the source-tract interaction in the glottal excitation model? In this section we discuss one possible approach--through the glottal area function.

The glottal area function measures the opening of the glottis during the production of voiced speech. Figure 35 shows a picture of the vocal folds; the area of the dark portion defines the glottal area function. Since the glottal area function is a measure of vocal fold mechanical movements, it is believed to be unaffected by the acoustic interaction between the glottal source and the vocal tract. Thus the glottal area function should remain the same over the course of a sentence. In this sense, a single glottal area function can be used to represent the glottal source characteristics over the course of a sentence.

The question is, how do we transform the glottal area function into the glottal volume velocity which includes the effect of source-tract interaction? This transformation is accomplished through an impedance circuit (Figure 36) proposed by Guerin et al. [15]. The circuit represents the input port of

Figure 35. Picture of a human vocal fold.

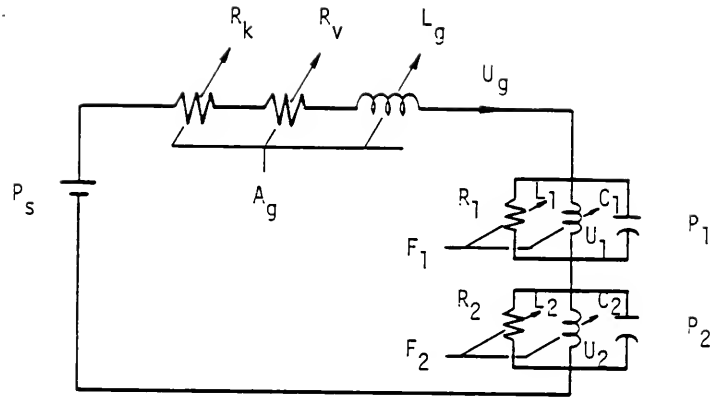


Figure 36. Impedance circuit for computing the glottal volume velocity.

the vocal tract filter shown in Figure 8. The vocal tract input impedance for each formant is modeled as a parallel RLC circuit. Since only the first two formant loading effects are significant, there are two RLC circuits in the impedance circuit. The glottal impedance is a function of glottal area (A_g) and glottal volume velocity (U_g). The experimental value for the glottal impedance has been obtained by van den Berg et al. [31]. The constant voltage source (P_s) represents the subglottal pressure. When the subglottal pressure, the formant frequencies, and the glottal area function are known, we can derive the glottal volume velocity (U_g) by solving the following set of differential equations.

$$P_s = U_g(t)\,[R_k + R_v] + L_g \frac{d}{dt}\, U_g(t) + P_1 + P_2$$

$$P_1 = L_1 \frac{d}{dt}\, U_1(t)$$

$$U_g(t) = U_1(t) + \frac{P_1}{R_1} + C_1 \frac{dP_1}{dt}$$

$$P_2 = L_2 \frac{d}{dt}\, U_2(t)$$

$$U_g(t) = U_2(t) + \frac{P_2}{R_2} + C_2 \frac{dP_2}{dt}$$

Since R_k, R_v, and L_g are functions of A_g and U_g, this is a set of nonlinear, time-varying differential equations. So the glottal volume velocity can only be solved by using numerical methods. We have used the Runge-Kutta method to solve the equations [32]. Figure 37 shows a comparison of the glottal area function and the derived glottal volume velocity. The most prominent distinction is that formant ripples are introduced to the glottal volume velocity. This phenomenon agrees with the observation of inverse filtering experiments.

Now that we realize that the glottal excitation function can be derived from the glottal area function, let us see how we obtain the glottal

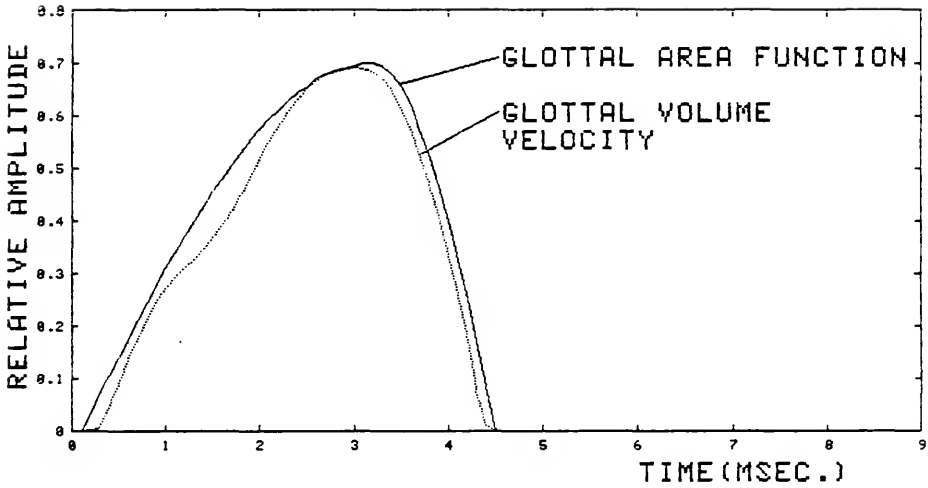


Figure 37. Comparison of the glottal area function and the glottal volume velocity derived from the glottal area function.

area function. There are two approaches for measuring the glottal area function. One approach is to measure the glottal area directly, the other is to estimate the glottal area function from the speech signal.

The technique we used to directly measure the glottal area is called ultra high speed cinematography [25]. The picture of the vocal folds, as shown in Figure 35, was taken at a frame rate of 5,000 frames per second. The glottal area for each frame is measured by an interactive computer image processing system. The glottal area as a function of frames is shown in Figure 38. The measured glottal area function is sampled 5,000 times per second, while the sampling rate of the speech signal is 10 KHz; thus a 1:2 interpolation has to be applied to the glottal area function to synchronize it with the speech signal. The disadvantage of the direct method is that it is an invasive method and cannot be applied to continuous speech, so we have to rely on the indirect method.

The indirect approach for obtaining the glottal area function is to estimate the glottal area function from the glottal volume velocity. This is the inverse of the transformation from the glottal area function to the glottal volume velocity.

Unfortunately, the transformation is a nonlinear process, so the inverse transformation cannot be carried out. An alternate way is minimum mean-square error fitting as illustrated in Figure 39. The glottal volume velocity, U'_g, for stylized glottal area function is computed and compared with the glottal volume velocity, U_g, obtained by inverse filtering the speech signal. The mean squared error between the spectra of the two normalized glottal volume velocities is computed as

$$E = \frac{1}{M} \sum_{m=1}^{M} U_g(m) - U'_g(m)$$

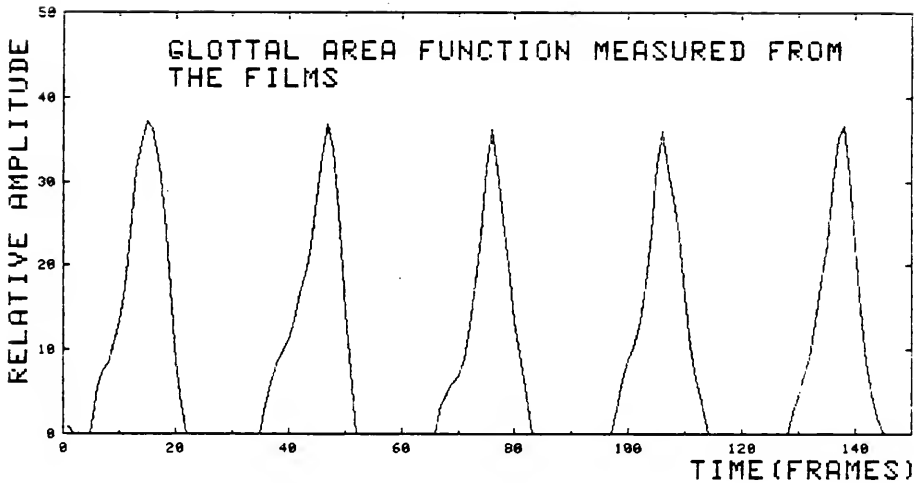


Figure 38. Glottal area function obtained by the high speed filming technique.

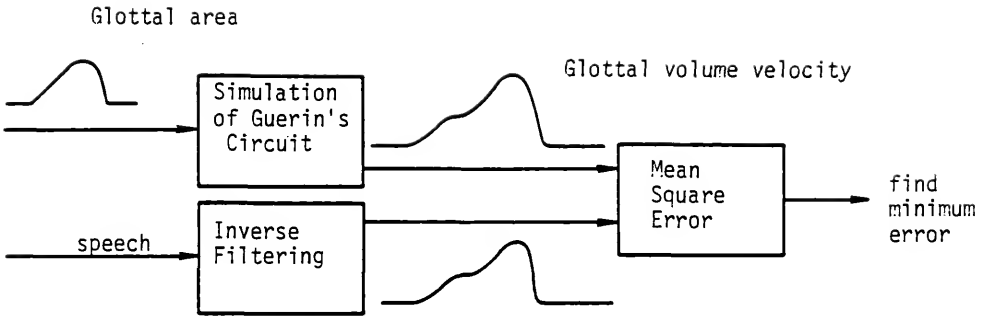


Figure 39. The procedure for estimating the glottal area function from speech.

where $U_g'(n)$ is the Discrete Fourier Transform (DFT) of the normalized volume velocity, $U_g'/E_{u'g}$, with $E_{u'g}$ being the energy of the waveform U_g'; similarly, $U_g(n)$ is the DFT of the normalized volume velocity, U_g/E_{ug}. The errors are calculated for a set of stylized glottal area functions which are obtained by varying the opening and closing phase of a model waveform shown in Figure 40. The stylized waveform corresponding to the minimum error is chosen as the best estimate of the glottal area function. Figure 41 shows a three dimensional plot of the inverted error function versus the opening phase and the closing phase of the glottal area function. The peak in this plot corresponds to the minimum error or the best estimate of the glottal area function. The bottom of Figure 41 shows the comparison of the inverse-filtered waveform and the model-generated waveform.

One last parameter we are going to discuss in this section is the subglottal pressure, P_s. The subglottal pressure is the driving force for the glottal air flow, as shown in Figure 36. During the production of a sentence, the subglottal pressure is not constant. We need to know the time variation of P_s for the simulation of Guerin's circuit. Previous studies [33,34,35] have shown that the subglottal pressure is related to both the intensity and the pitch of speech. For medium pitch phonation, the intensity of speech is roughly proportional to the 3.3 ± 0.7 power of the subglottal pressure. On the other hand, the pitch increases 2.5 Hz with a cm-H_2O increase in subglottal pressure. Thus, it seems that the subglottal pressure is a nonlinear function of both intensity and pitch of speech. Fortunately, it was found by the two mass model simulations of vocal cord motion that the fundamental frequency mainly depends on the vocal cord tension [36]. Thus, we adopted a simple rule for estimating the subglottal pressure from the sound intensity (I)

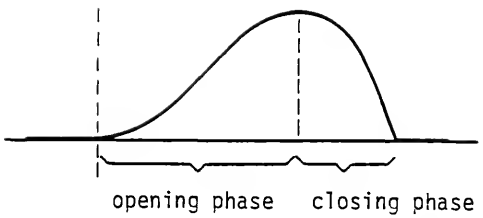
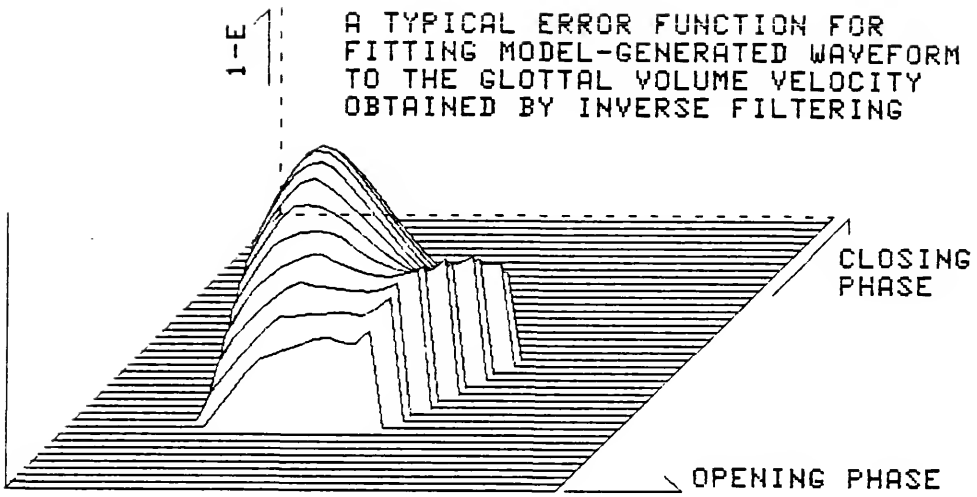


Figure 40. The model of glottal area function.



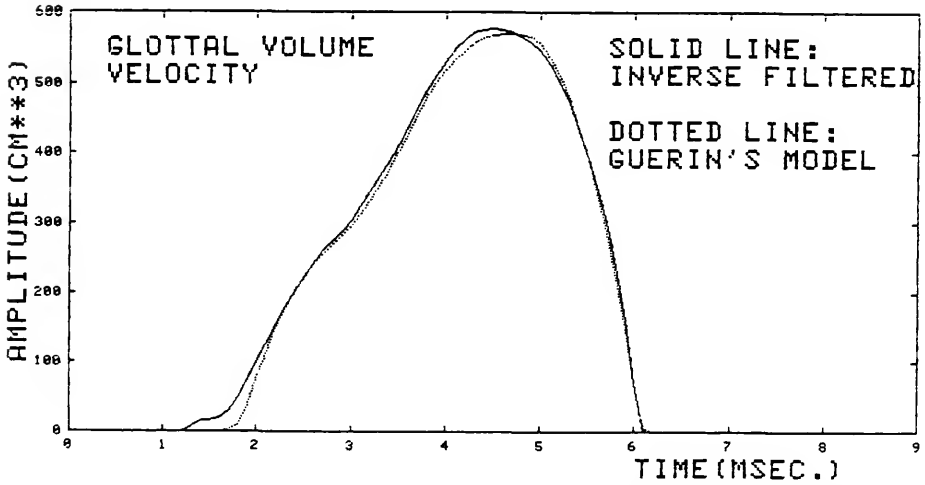


Figure 41. The three-dimensional plot of the inverted error function (top). Comparison of the inverse-filtered waveform and the model generated waveform (bottom).

$$P_s = 10^{\frac{I+C}{4}} \qquad (cm\text{-}H_2O)$$

where C is a constant. The validity of this relation is not proved theoretically, but it provides a workable relation and seems to produce a good result when synthesizing speech.

Synthesis of Sentences

The only way to show that one speech synthesis algorithm works better than others is by comparing the quality of the synthesized speech. There is a problem in evaluating the quality of vowels because the vowels are meaningless, while the listener usually tries to associate a meaning to what he listens to. Therefore, we decided to synthesize sentences for the purpose of speech quality evaluation.

We have synthesized three sentences for this research. One sentence is "We were away a year ago" by a female child. The second sentence is "The boy was there when the sun rose" by an adult female. The third sentence is "We were away a year ago" by an adult male. In this section we are going to discuss the details of synthesizing the above three sentences case by case. We will point out the important principles of speech synthesis involved. We will also discuss the difference between the child/adult speech and male/female speech.

Synthesis of the Child's Sentence

The child subject is STA. The sentence "We were away a year ago" was used as a test sentence in several previous speech analysis-synthesis experiments. This sentence is used widely in speech research since it is characterized by fast formant transitions which cause problems in both analysis and synthesis. Thus the success of any algorithm with this sentence is an indication of its usefulness.

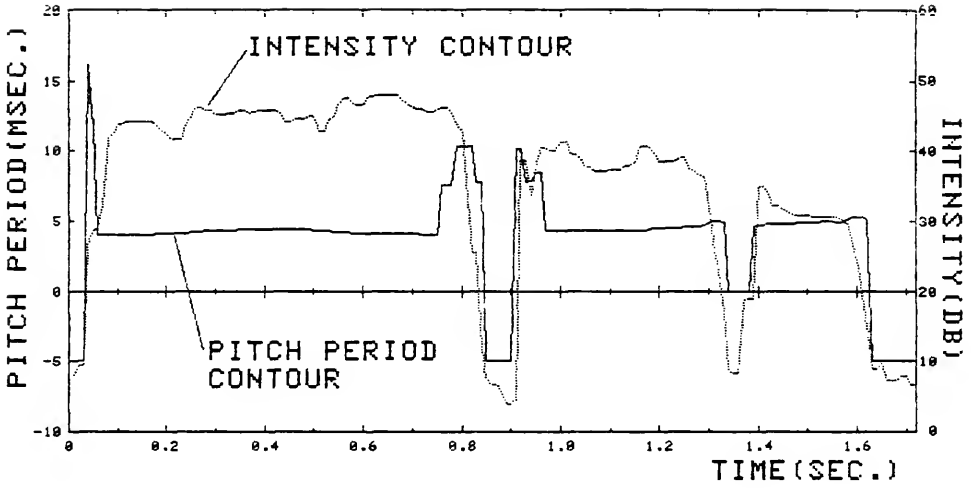
The most distinct feature of a child's speech is its high pitch or fundamental frequency. In this particular case, the fundamental frequency was 250 Hz (or the pitch period is 4 ms). This pitch period contour was extracted using the modified autocorrelation method (see Figure 42). Note the sudden jump in pitch period (called pitch break) both at the beginning and the end of voicing. This is caused by the switch of pitch control mechanisms from the subglottal pressure to vocal cord tension, and vice versa. The intensity contour is superimposed on the pitch period contour as a dotted line.

Another characteristic of the child's speech is that only four formants were found in the frequency range of interest, i.e., 0 - 5 KHz. This is due to the shorter vocal tract length for children; thus, the average frequency spacing between formants becomes larger, resulting in few formants. The formant contours for the entire sentence is shown in Figure 43. Notice that the formants (especially the second formant) are changing very rapidly due to the semivowels /w/, /r/, and /j/.

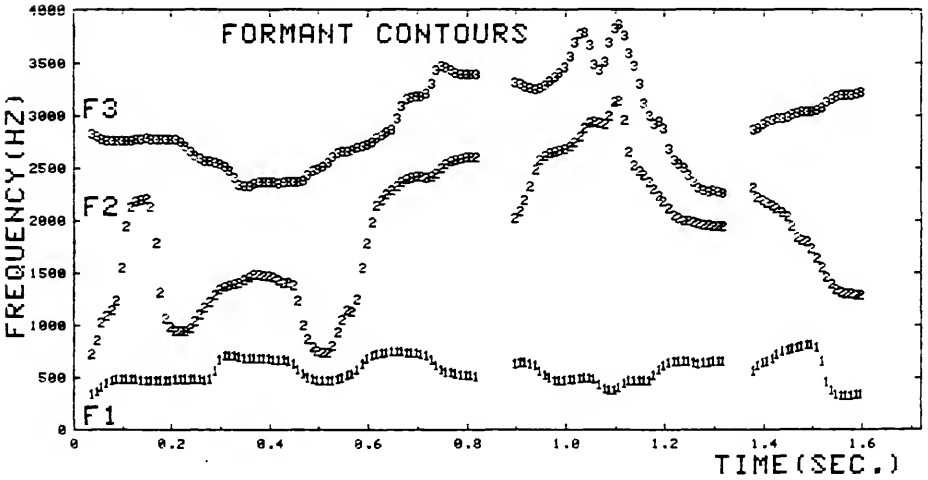
One problem we encountered in estimating the formant frequencies of the child's speech is that the first formant estimation is strongly influenced by the harmonics of the fundamental frequency. Because of the high-pitched characteristics of the child's speech, the error in the first formant frequency estimation can easily exceed 100 Hz (or 20% error). According to Flanagan [19] this amount of error will induce a noticeable change in the quality of speech, e.g., the just noticeable difference in formant frequency is 3 - 5 %. The effect of the fundamental frequency on the formant frequency estimation is illustrated in Figure 44. The figure shows both the short time spectrum (dotted line) and the linear prediction spectral envelope (solid line). Notice that there are two peaks in the spectral envelope

Figure 42. The pitch period contour and intensity contour of the sentence "We were away a year ago."

Figure 43. The formant contours of the sentence "We were away a year ago."



WE WERE AWAY A YEAR AGO



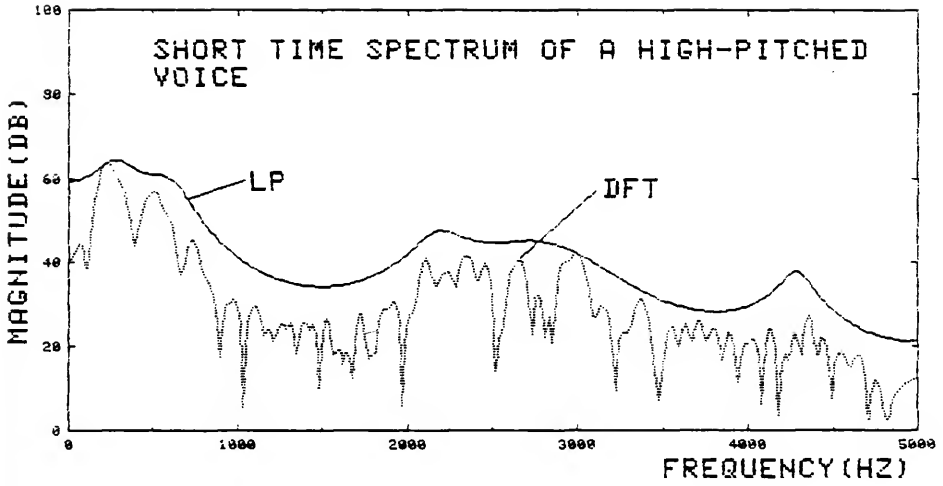


Figure 44. Spectrum of vowel /o/ for the child subject.

corresponding to the second and the third harmonics of the fundamental frequency. Either one may be labelled as the first formant, while the "real" first formant may lie somewhere between the two. The higher formants could also be in error. But the percentage of the error with respect to the formant frequency is usually less than 5% (the just noticeable difference) so it will not affect the quality of the synthetic speech.

A way to reduce the error in the first formant frequency estimate is to use the pitch synchronous linear prediction analysis. This reduces the influence of the pitch pulses on the estimation of the autocorrelation function, and hence increases the accuracy of the formant estimation.

The glottal volume velocity waveform can be obtained by inverse filtering the speech waveform. Figure 45 shows the glottal volume velocity waveform of vowel /i/ for the child's speech. Notice that the closed phase interval is very short in this case. This phenomenon is usually observed in children's speech and is associated with the breathy quality of the speech. The glottal volume velocity waveform can be used to construct the glottal excitation functions for the formant synthesizer as discussed in the previous section. The glottal excitation functions used to synthesize the child's speech are shown in Figure 46.

The parameters and excitation functions obtained by the analysis are used to resynthesize the speech. Since the sentence consists of non-nasal sounds throughout, no additional effort for setting the nasal formant/anti-formant is needed. The only thing we have to adjust in this sentence is the amplitude of frication and aspiration for the voiced plosive /g/ as in "ago". In this case, we set the amplitude of frication to 49 dB for 10 ms to give a sudden burst of noise, and then set the amplitude of aspiration to 30 dB for 10 ms before the voicing onset. This parameter

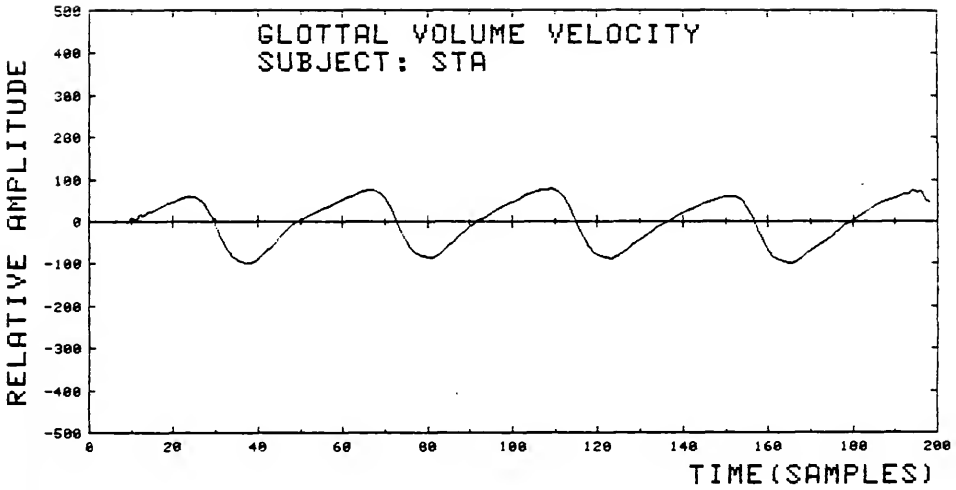


Figure 45. The glottal volume velocity waveform obtained by inverse filtering.

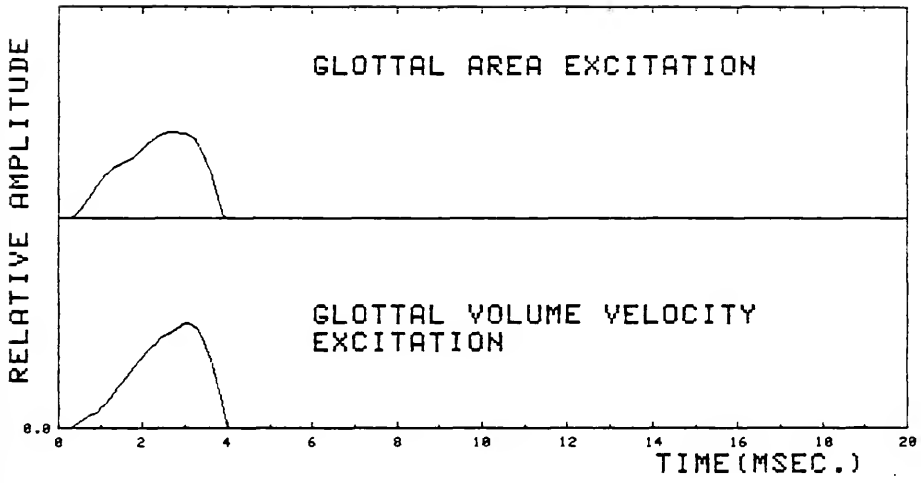


Figure 46. The glottal excitation functions used to synthesize the child's sentence.

setting agrees with the principle of synthesizing plosive sounds as discussed in the previous section. The time-domain characteristics of the synthetic consonant also agrees with the characteristics of the real speech, as shown in Figure 47. The quality (at least the intelligibility) for the synthetic speech is good compared to the natural speech. Figure 48 shows a comparison of the spectrograms of the real speech and one of the synthetic speech sentences, which used the glottal area excitation source. Notice that the high frequency portion is very noise, which accounts for the breathy quality of the voice.

Synthesis of the Female's Sentence

The female subject is DEB. The sentence "The boy was there when the sun rose" was extracted from the list of phonetically balanced sentences in IEEE Recommended Practice for Speech Quality Measurements [37]. As in the case of the sentence "We were away a year ago", this sentence is also characterized by fast formant transitions due to both semivowels and diphthongs. This sentence also contains the nasal consonant /n/ in two places, "when" and "sun." This makes the synthesis task a little more difficult because the model for nasalized sounds is, at best, an approximation.

The fundamental frequency range of this speaker is near 200 Hz, which is a typical fundamental frequency for female subjects. The pitch period contour for the sentence is shown in Figure 49. Unlike the case of the child's speech, there is no "pitch break" for our female speech example. This probably indicates that the adult female speaker had more control of her voice than the child did. The intensity contour of the sentence is superimposed on the pitch period contour as a dotted line (see Figure 49).

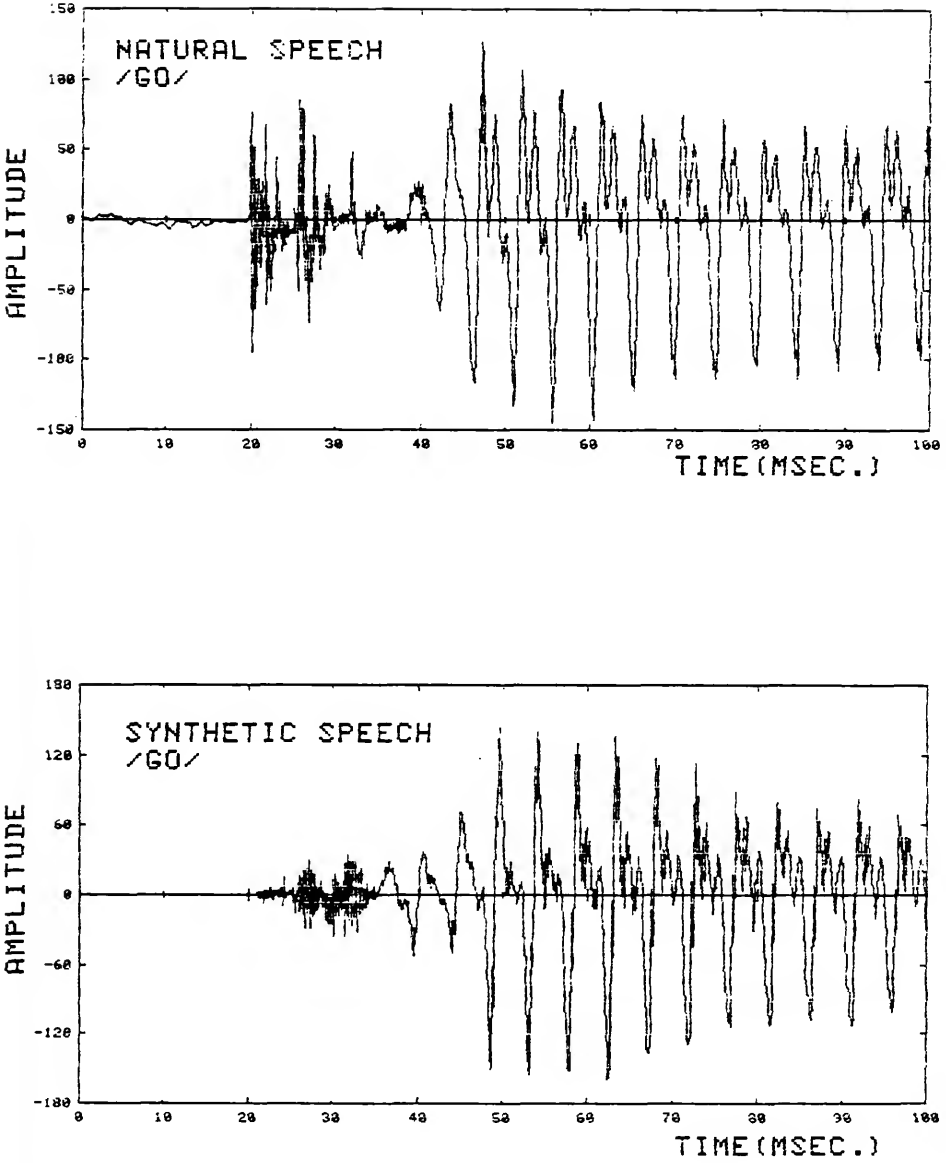


Figure 47. Time waveforms of the natural and synthetic plosive consonant /ɑ/.

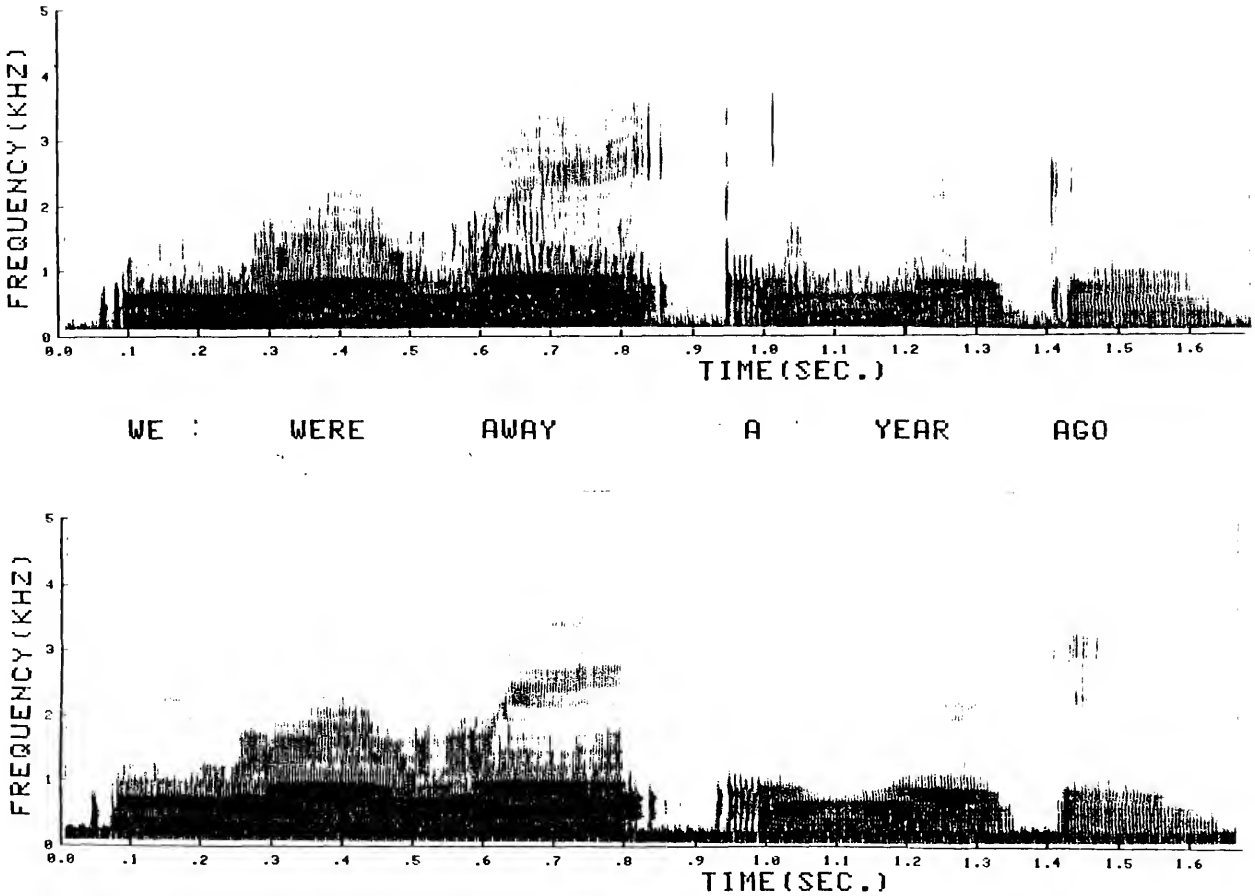


Figure 48. Spectrograms of natural (top) and synthesized (bottom) child's sentence.

The first three formant frequency contours are shown in Figure 50. Notice that the sentence is characterized by rapid formant transitions, as mentioned before. Another feature of this feminine speech example is that it has four formants in the frequency range of interest, i.e., 0 - 5 KHz. This can be seen in the short time spectrum of vowel /o/ in Figure 51. In formant analysis, we found it difficult to extract the formant frequencies of nasal vowels and nasal consonants. The effects of nasalization are discussed in previous papers [38,39]. The characteristics of nasal sounds are illustrated by the spectrum of the nasal consonant /n/, as shown in Figure 52.

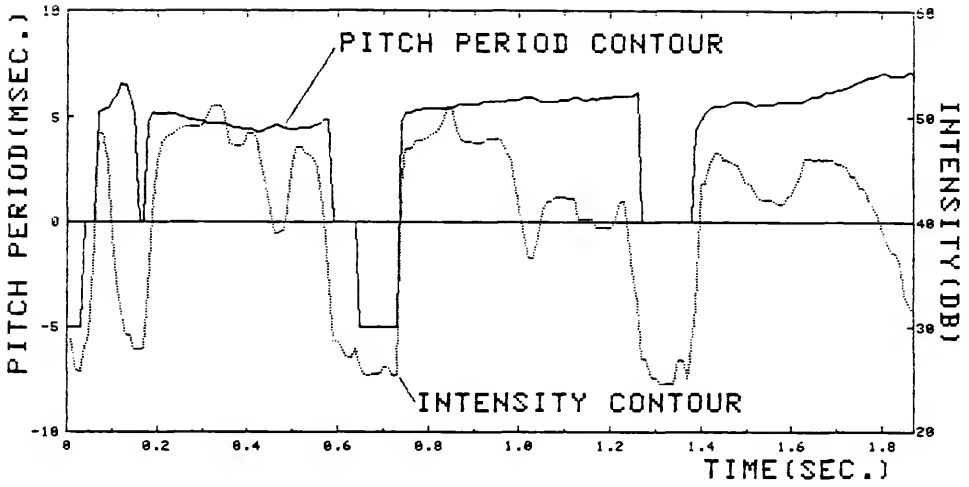
Two features of nasalized sounds have been observed in our experiments. The first is the split formant phenomenon. Due to the coupling of the nasal tract, an additional formant/antiformant pair is introduced in the transfer function of the vocal tract filter. The effect is to broaden the first formant bandwidth in the nasalized vowel, or to cancel the first formant in the case of nasal consonants. The second feature is that the higher formant peaks are obscure due to the losses of the nasal tract.

Thus, the vocal tract transfer function is characterized by both poles and zeros. Unfortunately, there are no effective methods of modeling a pole-zero system.

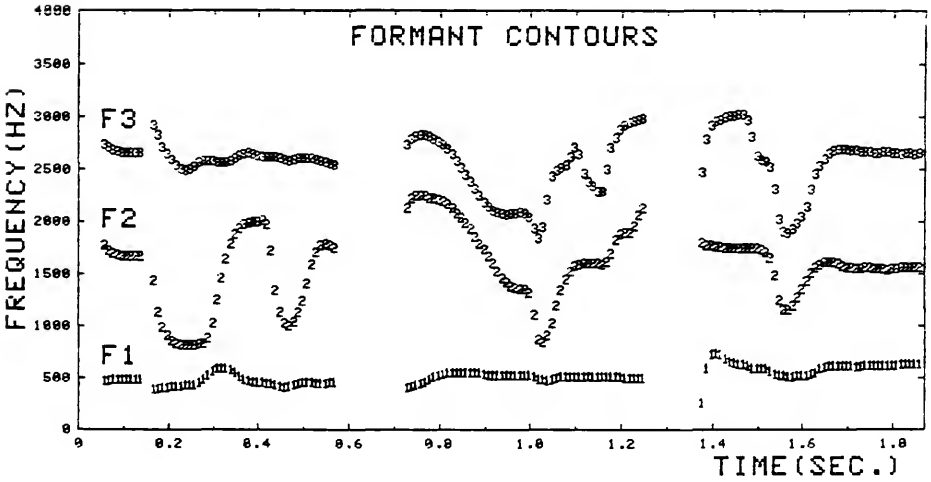
We have to use the LP analysis, which is suitable for analyzing an all-pole system, to analyze nasal sounds. This, of course, resulted in inaccurate formant estimation. Also, the nasal antiformant frequency is not easily estimated. However, we have developed an algorithm for detecting the nasal sound. We can then adjust the nasal antiformant frequency in the nasal sounds to obtain good quality speech. The details of the algorithm are discussed in Appendix B.

Figure 49. The pitch period contour and the intensity contour for the sentence "The boy was there when the sun rose."

Figure 50. The formant contours of the sentence "The boy was there when the sun rose."



THE BOY WAS THERE WHEN THE SUN ROSE



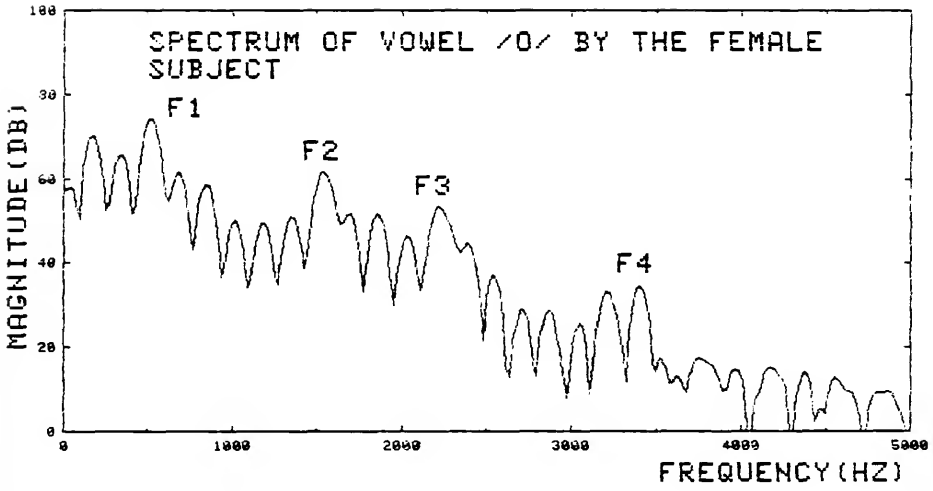


Figure 51. Spectrum of vowel /o/ by the female subject.

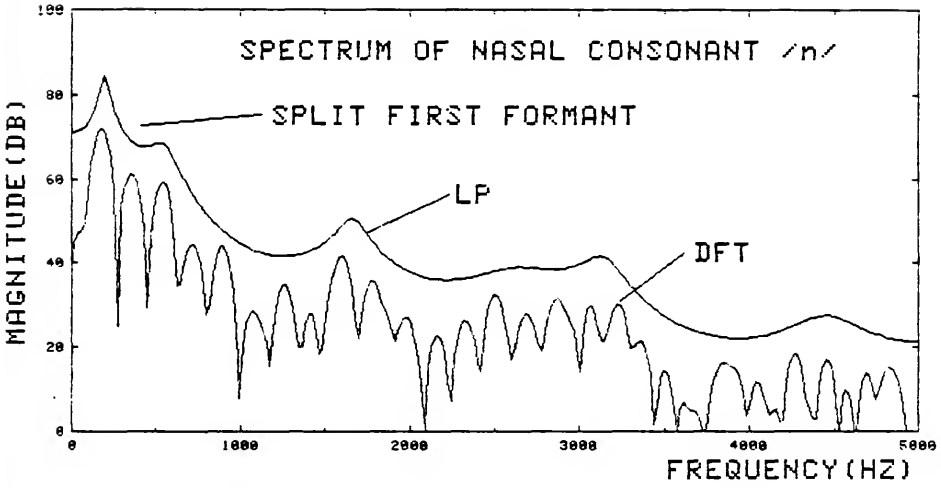


Figure 52. Spectrum of the nasal consonant /n/.

The glottal volume velocity waveform is obtained by inverse filtering the speech. A typical volume velocity waveform is shown in Figure 53 for vowel /o/ as in "rose." We can see that the closed phase period (flat portion) is longer than the child's case. Thus the closed-phase analysis can be accomplished. The inverse-filtered waveform can be used to construct the glottal excitation functions as shown in Figure 54.

The parameters and the excitation functions are used to resynthesize the sentence. As expected, the difficulty in synthesizing the sentence is related to the nasalized sounds. In Klatt's formant synthesizer, nasalized sounds are synthesized by inserting a formant/antiformant pair in the vocal tract transfer function. In the case of nasal consonants, the antiformant frequency should be very close to the first formant frequency so as to cancel the first formant. Thus there is only a dominant nasal formant peak in the first formant region, which can be estimated accurately by the LP analysis. So, nasal consonants can be synthesized very easily. The difficult problem is to estimate the antiformant frequency for the nasalized vowels, since the antiformant may be located anywhere between the nasal formant frequency and the first formant frequency depending on the degree of nasalization. There are no accurate methods for estimating this antiformant frequency, so we had to use a trial-and-error approach to obtain an optimal antiformant frequency contour. After a few trials, the nasalized vowels were made quite natural. There are still some imperfections due to the improper modeling of nasalized vowels, but our main interest is in the glottal excitation function; therefore, we did not pursue this problem further. Figure 55 shows the spectrograms of the real speech and the synthetic speech sentence with the glottal area as the source of excitation. We can see that the formant contours were accurately reproduced in the synthetic speech.

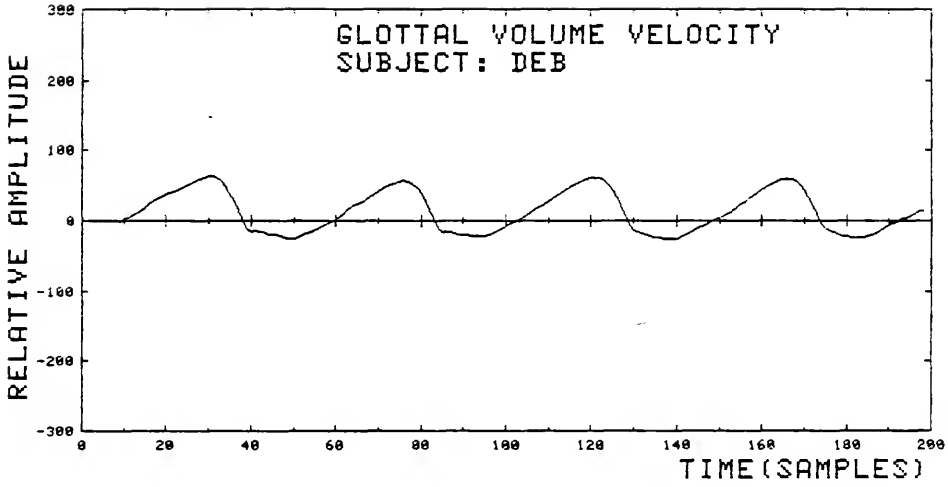


Figure 53. The glottal volume velocity waveform obtained by inverse filtering.

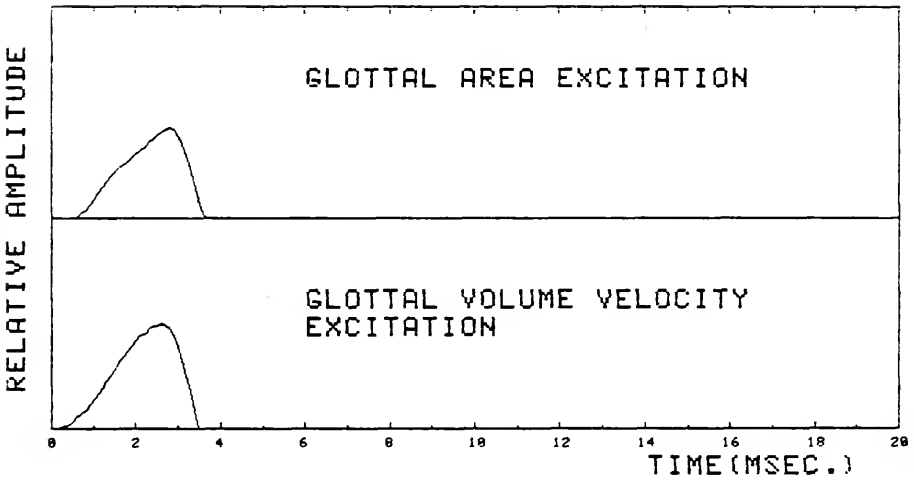


Figure 54. The glottal excitation functions used for synthesizing the female's sentence.

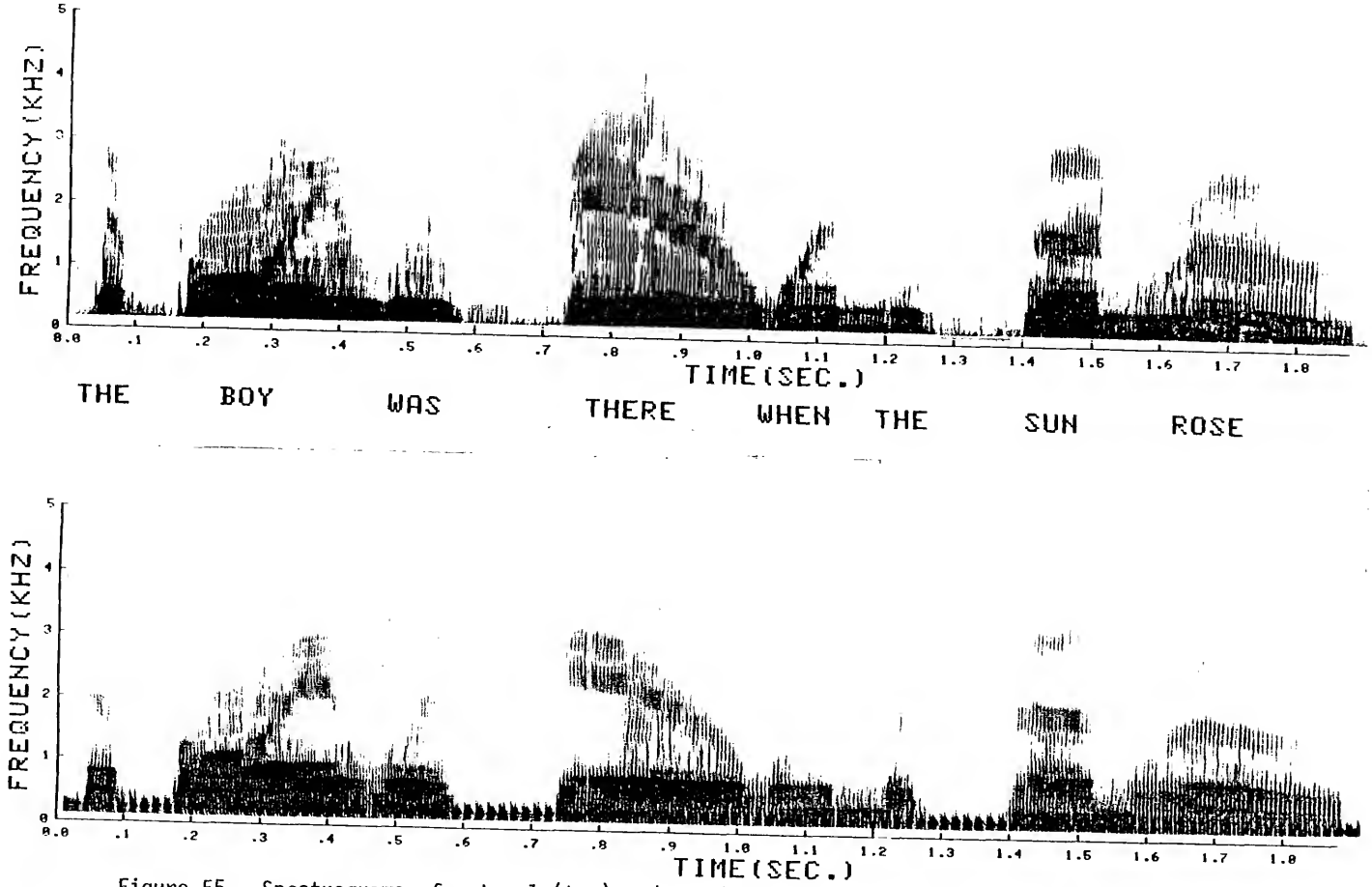


Figure 55. Spectrograms of natural (top) and synthesized female's sentence (bottom).

Synthesis of the Male's Sentence

The subject in this case is DGC. The sentence is again "We were away a year ago." The reason for using this sentence is the same as for the child subject. Another characteristic of the sentence is that the speech is voiced and not nasalized throughout the sentence. This makes the analysis and synthesis of this sentence easier than the previous female's sentence.

The pitch period contour of the sentence is shown in Figure 56. The average pitch period of this speaker is 7.5 ms, which corresponds to a fundamental frequency of 130 Hz. Notice that the speech is identified as voiced throughout the sentence, even for the voiced fricative /g/. This is different from the case of the child's sentence where /g/ is labelled as unvoiced. Thus, we have to add a frication source in synthesizing this phoneme. The intensity contour is also shown in Figure 56 as the dotted line.

As we discussed in Chapter 2, the male's speech is characterized by five formants in the frequency range up to 5 KHz. This is verified by the spectrum of vowel /i/ in Figure 57. The first three formant contours are shown in Figure 56. The formant contours look very similar to the formant contours for the child's sentence, except that the duration of each phoneme may be different in the two cases.

The glottal volume velocity waveform of the vowel /e/ is shown in Figure 58. One characteristic of this waveform is that the closed-phase period is very short; most of the time we cannot identify a clear closed-phase period. This is a characteristic of medium or low intensity speech. Another thing worth noting is that the formant ripples are more prominent in this case than in the cases of feminine and child speech. A possible

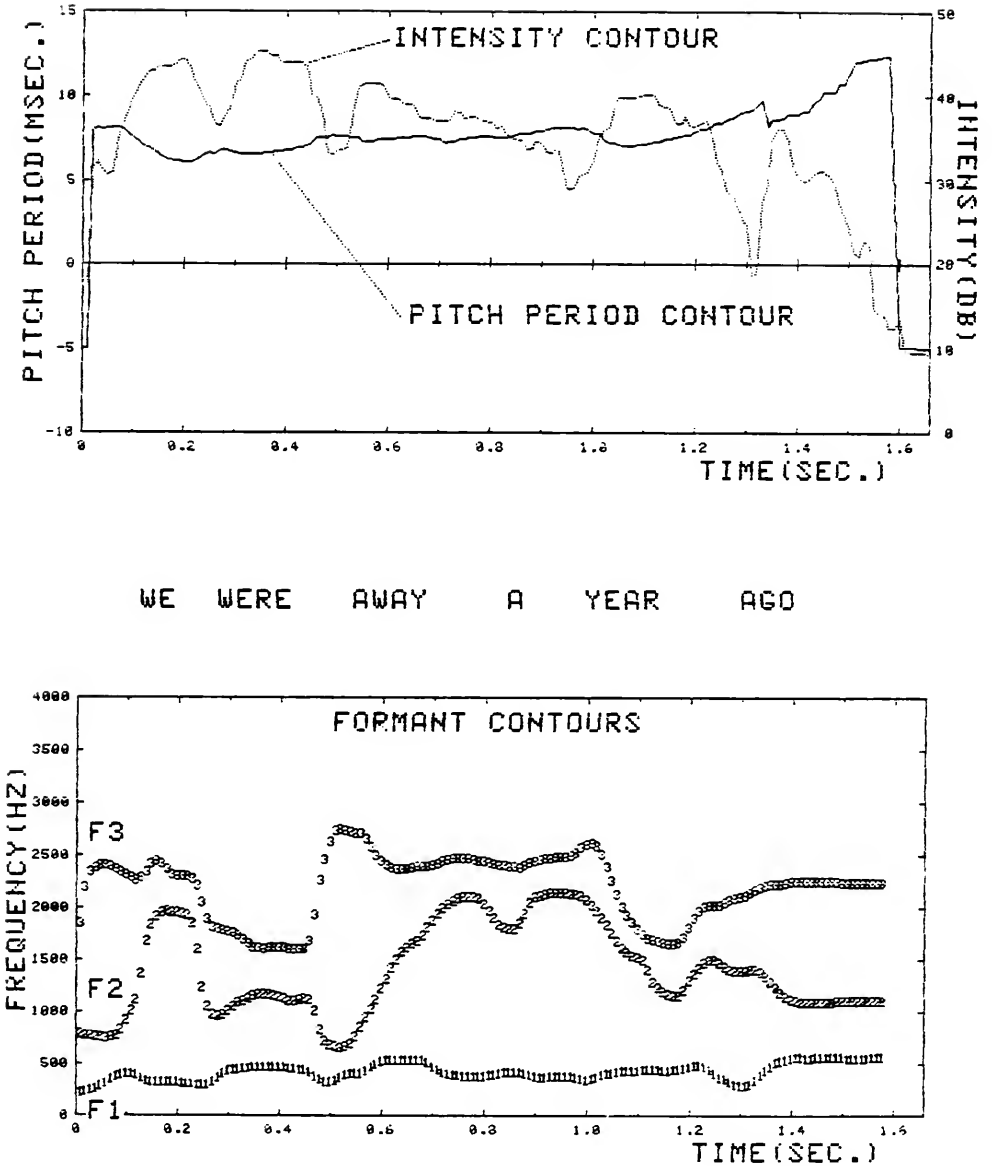


Figure 56. The pitch period contour and the intensity contour (top) of the sentence "We were away a year ago." The formant contours of this sentence are shown in the bottom section.

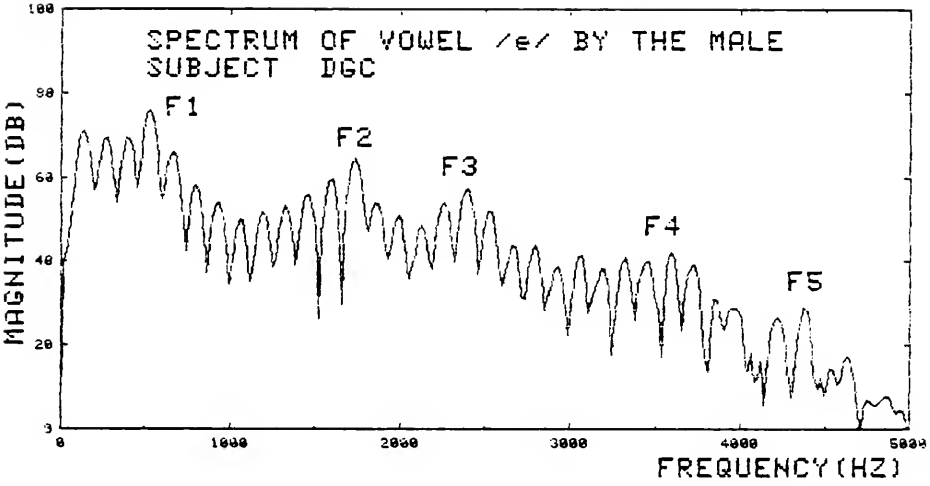


Figure 57. The spectrum of vowel /e/ for the male's speech.

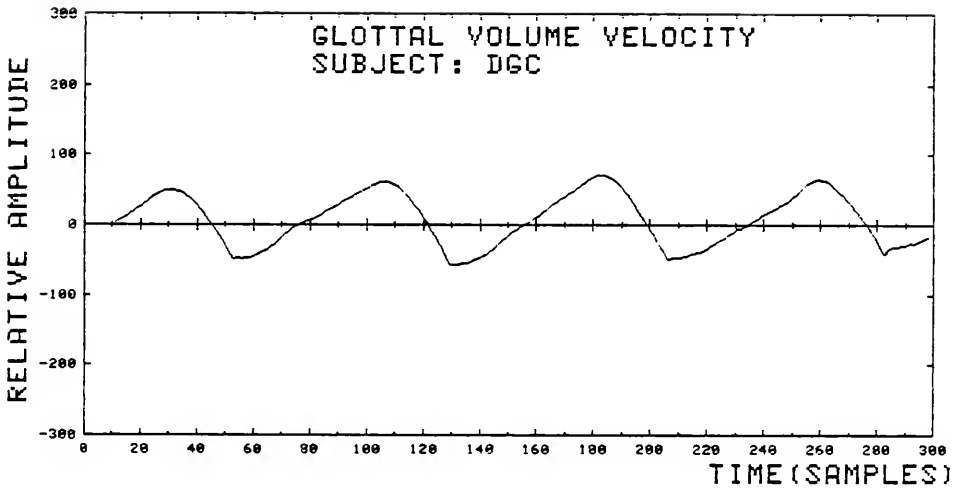


Figure 58. The glottal volume velocity waveform obtained by inverse filtering.

explanation is that the size of the glottal opening is larger for the adult male subject, which reduces the glottal impedance and hence increases the loading of the vocal tract filter. We have taken care of this difference by using a smaller glottal impedance in synthesizing the male's speech.

The glottal excitation functions used for synthesizing this sentence are shown in Figure 59. As expected, the synthesis of this sentence is easier than the female sentence, because it does not contain nasalized sounds. Figure 60 shows the spectrograms of the real sentence and the synthesized sentence (with the glottal area excitation). The quality of synthetic sentences with different excitation functions are quite different in this case because 1) the effect of source-tract interaction is more significant, and 2) the "buzziness" of synthetic speech is more obvious in the low pitch male voice. We will discuss more on the quality of speech in the next chapter.

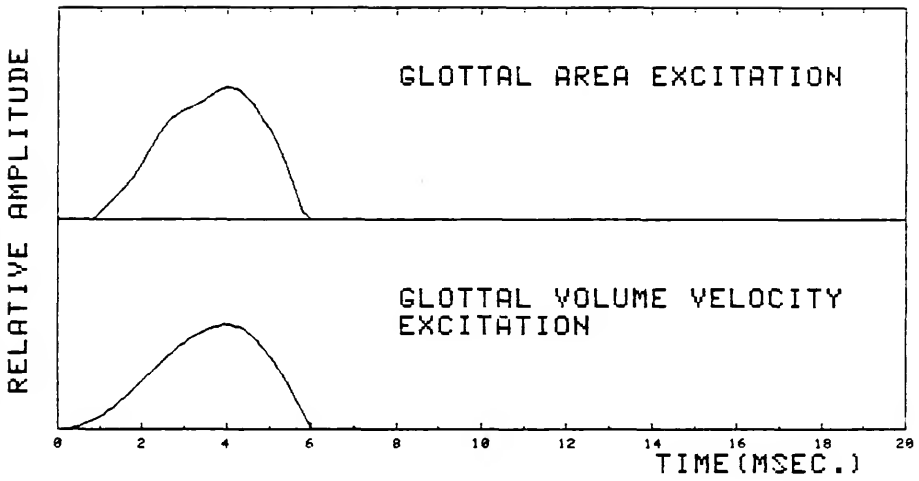


Figure 59. The glottal excitation functions used for synthesizing the male's sentence.

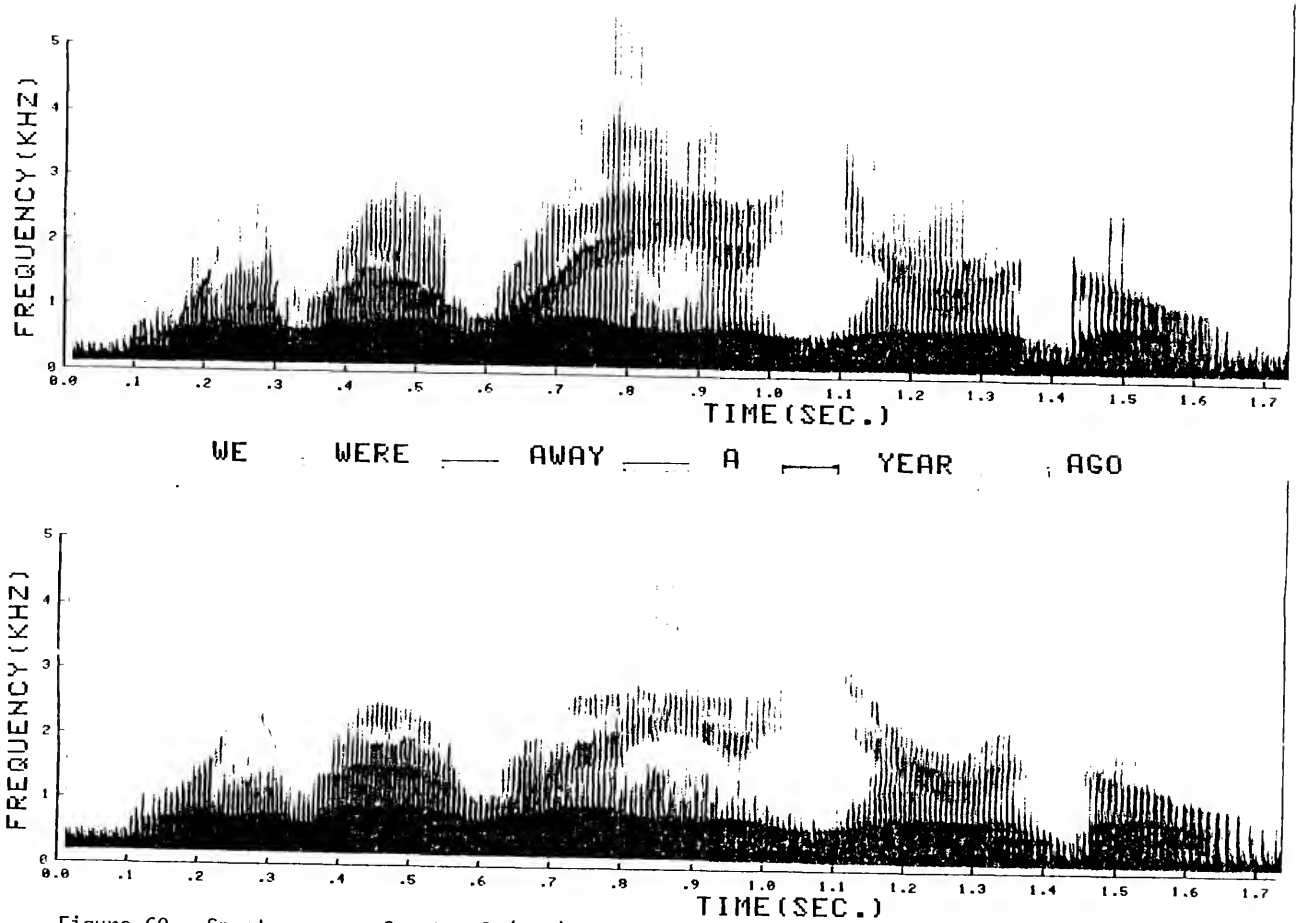


Figure 60. Spectrograms of natural (top) and synthesized (bottom) male's sentence.

CHAPTER 4
EVALUATION OF THE QUALITY OF SYNTHETIC SPEECH

The goal of this research is to investigate which type of glottal excitation function is the best for synthesizing high quality speech. In order to achieve this goal, we need to have a good method for evaluating the quality of speech. There are both objective and subjective methods for evaluating the quality of speech. From the engineering point of view, it is preferrable to have an objective method because the results obtained by the objective method are reproducible. But the objective methods developed so far only measure the "intelligibility" of speech [40]. Since we are interested in the overall quality of speech, we have to use the subjective method of speech quality measurement. An argument in favor of the subjective method is that speech quality is a "subjective measure" itself. In this chapter we are going to discuss the definition of "speech quality" and the subjective methods of speech quality measurement. We will design a formal listening test to evaluate the quality of synthetic speech discussed in Chapter 3. And we will discuss the results of the listening test.

The Concept of Speech Quality

A study of speech quality measurement procedures requires a working definition of the overall concept of speech quality. The concept of speech quality encompasses the total auditory impression of speech on a listener and not just its intelligibility aspect. Speech quality includes additional factors such as loudness, naturalness and clarity, speaker identifiability,

timbre and rhythmic character, systematic amplitude or time distortions, and many others. The relative importance or even a closer definition of all these factors will be influenced by characteristics of the system to be evaluated--e.g., the degradation or loss of "quality" in transmitting speech over a telephone system may be seen as completely different from the degradations in a vocoder system or other speech synthesis system. In the case of the telephone system, speech quality includes the vocal characteristics of the speaker. Here, "naturalness" of the signal would be a measure for the ease of identifying the original speaker. In the other case, the naturalness of a "synthetic voice" describes how human the speech signal sounds and--because of the anonymity of such an artificial voice--no relation to a particular speaker exists. For such reasons, a quantitative definition of the different factors is often not only difficult but sometimes next to impossible.

In this study we adopt a simple definition of speech quality offered by Rothauser et al. [39,41]. According to this definition, speech quality is described in terms of only four parameters: intelligibility, preference, loudness, and speaker recognizability. With the exception of loudness, none of the four parameters is firmly defined by a single generally recognized measurement procedure so that there is adequate freedom for the establishment of proper definitions. We will not go into the problem of parameter definitions. Instead we will concentrate on the parameter "preference." This term describes the average attitude of a listener toward a speech signal while he is comparing it consecutively with another speech signal. Preference thus provides an answer to the question: Which one of the two speech signals to be compared is preferred by an average listener as a source of information? The aspect of preference with respect to the

overall speech quality becomes dominant when all the following conditions are fulfilled, which may often be true for practical cases:

a) The intelligibility of the speech signal is high enough so that it loses its importance as a prime quantitative speech quality parameter and design criterion.

b) The level of the speech signal is kept in a setting of optimum loudness, a condition that eliminates the influence of loudness on the quality of the speech signal to be evaluated. Optimum loudness is given by the average speech level at which the single listener or a listener group prefers the sound level of the speech presented.

c) The recognizability of the speaker should be of no interest to the listener. This is the case for all systems where the listener does not expect more information from the speech signal than he could get from written text.

Under these circumstances, preference may be said to represent speech quality for all practical purposes.

In our study, all three of the above conditions are satisfied because a) the intelligibility of the synthetic speech is high and independent of the glottal excitation function, b) the loudness of the speech can be adjusted to provide the most comfortable listening condition, and c) the speaker identity is of no interest to the listeners since they do not know the speaker. Thus preference represents the quality of speech. More specifically, preference represents the "naturalness" of speech, since the intelligibility is of no consequence in this case.

Design of the Listening Test

Now that we understand the concept of speech quality, we can turn our attention to the design of the listening test. There are several important considerations in designing a listening test, such as the listener choice, the test format, the listening conditions, etc. We will discuss these factors in the following paragraphs.

The objective of this listening test is to measure the quality of different synthetic speech signals which we obtained by the procedures described in Chapter 3. These synthetic speech signals are called the "test signals." We will compare these test signals with the recorded natural sentence, which is called the "reference signal." We will also compare the quality between pairs of the test signals. In this way, we are adopting a pair-comparison test.

For each sentence, four signals are used in the listening test, including three test signals and one reference signal. The four signals are paired with each other to give six pairs. Each pair can be arranged in two different orders, i.e., (A,B) and (B,A). This will avoid the problem that listeners may favor the first signal or the second signal. Each pair will appear twice in the listening test. So, the total number of pairs is 24.

The pairs of speech signals are presented to the listeners in random order. Each pair is presented twice to form a group. The time relation of the speech signals is shown in Figure 61. The time interval between any two signals in a group is 500 ms. The time interval between two consecutive groups is 5 seconds to reduce the interference from the neighboring groups. This will also allow the listeners to have enough time to make their decision. We also put a 200 ms tone signal (1000 Hz) before each

TIMING OF THE LISTENING TEST MATERIAL

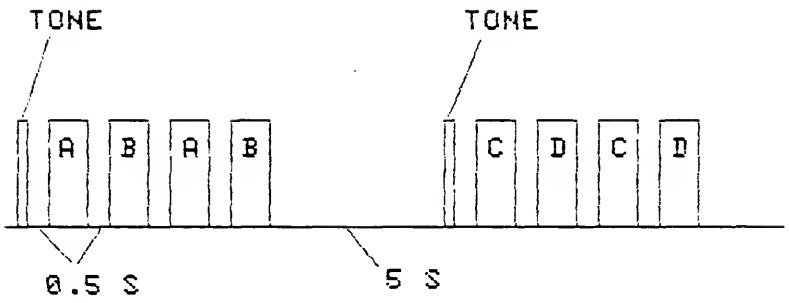


Figure 61. Timing diagram of the listening test materials.

group of speech signals to cue the listeners the arrival of the speech signals. The test materials are recorded on a REVOX tape recorder before presenting to the listener group.

The listener group consists of college students that are majoring in speech communications. Most of them do not have experience with synthetic speech. This group fits our purpose well because they have an adequate background in speech, which enables them to detect the anomalies of speech. On the other hand, they are not familiar with synthetic speech so they will not particularly concentrate on certain aspects of synthetic speech.

The listening test was held in a quiet room environment. The tape was played through a loud speaker. The loudness of the speech signal is adjusted to a comfortable level for all the listeners. One argument for using the loud speaker instead of the earphones is that the listening condition is more natural this way. However, we also used the earphones for a specific group of listeners, as we will discuss later.

The listeners compared each pair of speech signals in response to the question: "Which one of the two sounds was more natural?" They are instructed to mark the one that sounds more natural on an anwser sheet. There should not be a tie situation, i.e., both marked or not marked. So, this test is a "forced-choice" comparison.

There is a consideration about the listener's ability to make a good judgment about the quality of speech. The procedure of choosing the listeners who can make valid judgments is called "listener screening." This is usually done by deciding the consistency of the listener in making the required judgment. In this test, the consistency of the listener is decided by looking at his or her responses to two occurrences of the same pair of speech signals, i.e., (A,B) and (B,A). If the responses are consistent more than 75%

of the time, the listener is accepted. In this way, there is no need for a separate "listener screening test."

Another way of ensuring the good judgment of the listener is to form a group of experienced speech scientists. With this thinking, we asked two qualified persons to take the listening test. One of them, G. P. Moore, is a professor in the speech department at the University of Florida. He specializes in pathological voice detection. Another listener was D. G. Childers, a professor in electrical engineering at the same university, who has been engaged in speech research for ten years. To make the listening condition more critical, we asked the listeners in this group to wear earphones. Otherwise, the same tape and test conditions were used in this group as in the previous group. The data obtained from this listener group were analyzed independently from the previous group's data. We will discuss the significant difference between the two in the next section.

Results and Discussion

The listening tests for the two groups of listeners were held under different conditions. For the first listener group, the listening test was held in a sound-treated room. For the second listener group, the listening test was held in a quiet office environment, since headphone was used. The listening environments for the two listening tests are described in more detail in Appendix C. In this section we will summarize the results of the two listening tests and discuss possible differences.

Analysis of the results of the listening tests is facilitated by a simple scoring system. Each signal judged "more natural" is given a score of one, otherwise it is given a score of zero. The score given to a speech signal by a listener is then proportional to the subjective quality of the speech

signal. The highest score a speech signal can get is 12, corresponding to the most natural-sounding speech. The lowest score is zero which corresponds to the least preferred speech signal. If we average the scores over the listener group, we obtain a score which represents the average attitude of the listeners toward a particular speech signal.

Response of Listener Group One

The listeners in this group are master students in an audiology class. There are ten of them attending the listening test. Five of them did not pass the listener screening test, because their responses are inconsistent more than 25% of times. This high percentage of inconsistency is probably due to the high quality of synthetic speech. The responses of the rest five listeners are summarized in Table 5.

The quality of the speech signals is ranked in the following order, averagely.

1) Natural speech.

2) Glottal area function synthesized speech.

3) Glottal volume velocity synthesized speech.

4) Impulse excitation synthesized speech.

The rank order of the child's sentence is the same, while the ranking for the adult male and female sentences are a little bit different from the average. However, in both cases, the differences are too small to affect the overall ranking.

The results of the listening test can be presented in different ways. Table 6 summarizes the results of listening in a pair-comparison manner. The numbers represent the percentage of time that the sentences in the left column are judged more natural than the sentences in the upper row. This table is consistent with Table 5 in ranking the quality of different speech sentences.

Table 5. Average Scores* for the Speech Signals.

Subjects / Excitations	Child	Male	Female	Average
Natural	10.6	8.8	11.8	10.4
Glottal Area	6.2	5.4	5.4	5.7
Glottal Volume Velocity	4.6	5.8	3.2	4.5
Impulse Excitation	2.8	3.6	3.6	3.3

* Highest score is 12

Table 6. Pair-comparison Results Showing the Percentage of Times Sentence A Sounds More Natural than Sentence B.

A \ B	Glottal Area	Glottal Volume Velocity	Impulse Excitation
Natural	83%	87%	90%
Glottal Area		60%	73%
Glottal Volume Velocity			58%

The above results are for listeners with responses consistent more than 75% of the time. If we increase the consistency requirement to 80%, then only three listeners can pass the listener screening test. The responses of these three listeners (Tables 7 and 8) are consistent with the results for the five listener case. The only difference is that the scores are more widely separated from each other. This is a direct result of improved consistency in listeners' response.

On the other hand, if we average the scores over all listeners (without any screening), the rank order is still the same as in the previous two cases. The average scores are shown in Table 9. The only difference is that the scores for the three synthetic speech signals are closer in this case.

Results for Listener Group Two

The second listener group consisted of two professors. The difference between this listening test and the previous one was that

1) The listeners were considered qualified for making speech quality judgments, so no listener screening was needed.
2) The speech material was played through headphones so the listeners were able to make more critical judgments.

The significance of these two factors is that the listeners were very consistent in their responses. One of the listeners even achieved 100% consistency in his response. The results of this listening test are listed in Table 10 and 11 in two different manners.

Comparing the results obtained from this group with that of the previous group, we see that the ranking of speech quality is still the same. The difference is that the scores are more widely separated in this case. This is, again, a consequence of the consistency in listeners' response.

Table 7. Scores of the Listeners with 80% Consistency in Response.

Subjects / Excitations	Child	Male	Female	Average
Natural	11.7	9.7	11.7	11
Glottal Area	7.0	5.7	6.3	6.3
Glottal Volume Velocity	3.7	5.7	3.0	4.1
Impulse Excitation	1.7	3.0	3.0	2.6

Table 8. Pair-comparison Results for Listeners with 80% Consistency.

A \ B	Glottal Area	Glottal Volume Velocity	Impulse Excitation
Natural	89%	92%	94%
Glottal Area		75%	72%
Glottal Volume Velocity			69%

Table 9. Average Scores Over All Listeners

Excitations	Scores
Natural	9.6
Glottal Area	5.6
Glottal Volume	5.0
Impulse	3.7

Table 10. Scores Obtained from the Second Listener Group

Subjects / Excitations	Child	Male	Female	Average
Natural	12.0	11.5	12.0	11.8
Glottal Area	7.0	7.0	6.5	6.8
Glottal Volume Velocity	4.0	5.5	4.5	4.7
Impulse Excitation	1.0	0.0	1.0	0.7

Table 11. Pair-comparison Results for Listener Group Two Showing the Percentage of Times Sentence A Sounds More Natural than Sentence B.

A \ B	Glottal Area	Glottal Volume Velocity	Impulse Excitation
Natural	96%	100%	100%
Glottal Area		75%	92%
Glottal Volume Velocity			92%

In conclusion, the results of the listening tests showed that the quality of speech signals is consistently ranked in the following order:

1) Natural speech.

2) Glottal area excitation synthesized speech.

3) Glottal volume velocity synthesized speech.

4) Impulse excitation synthesized speech.

CHAPTER 5
CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

In this research we studied the effect of glottal excitation on the quality of speech. Three excitation functions were compared, namely, glottal area excitation, glottal volume velocity excitation, and impulse excitation. The common feature of these three excitation functions is that the spectrum envelope drops at a rate of -12 dB/octave. The unique feature of the glottal volume velocity excitation is that it has the general waveshape of a true glottal excitation but does not possess formant ripples. The unique feature of the glottal area excitation is that it can follow the detailed time characteristics of the true glottal excitation, including the formant ripples.

We used the three glottal excitation functions to synthesize sentences. The quality of the synthesized sentences were evaluated in a formal listening test using the natural sentences as references. The results of the listening test indicate that the quality of speech ranked in the following order:

1) Natural speech
2) Glottal area function synthesized speech
3) Glottal volume velocity synthesized speech
4) Impulse excitation synthesized speech

The results of the listening test can be interpreted as

1) The time-domain characteristics of the glottal excitation are more important than the spectral characteristics for the quality of speech.

2) The effect of source-tract interaction is important for generating high quality speech.

The implication of these two conclusions is that we should incorporate the time characteristics of the glottal excitation into a speech synthesizer. For efficient coding in formant vocoder applications, the glottal excitation can be represented by a 3-parameter model as shown in Figure 34. The parameters of the glottal excitation model can be obtained by inverse filtering the speech signal.

For very high quality speech synthesis, the glottal excitation model should include the effect of source-tract interaction. This interaction can be simulated using an impedance circuit proposed by Guerin [15] as discussed in Chapter 3, or the circuit proposed by Ananthapadmanabha and Fant [42]. Since the control parameters for this circuit are formants, intensity, and the glottal area, only the glottal area function need be added to the parameters already used for a common formant synthesizer. This is very attractive in telecommunication applications where the channel capacity is limited.

In order to apply the proposed formant synthesis technique to formant vocoder applications, some of the analysis procedures discussed in Chapter 2 need to be refined as follows.

1) The formant tracking algorithm should be automated. This can be done by using both the pitch/voicing contour and the intensity contour to segment speech into sections; that is, to use the peaks in the intensity contour as the "anchor" points, and use the valleys in the intensity contour or the unvoiced/silence region as the section boundary points. An illustration of this scheme is shown in Figure 62, where the intensity contour and the pitch/voicing

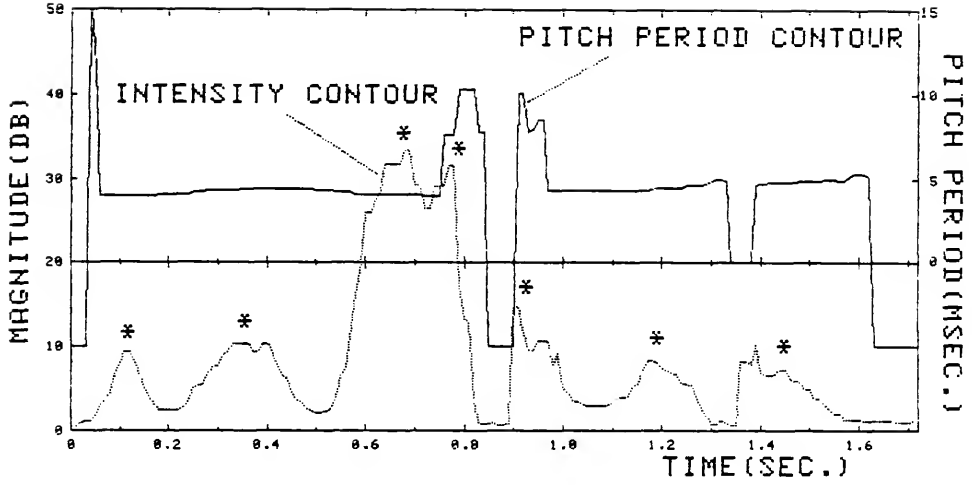
contour of "We were away a year ago" are shown. We observed that the peaks in the intensity contour can be used as the "anchor" points as marked in the figure. Notice, however, the intensity contour here is not the energy of speech; rather, it is the energy of the error signal in the LP analysis. The reason for not using the speech energy is that it is characterized by undulations in the curve (Figure 63), so a simple peak picking algorithm will not work.

2) A pitch-synchronous or closed-phase analysis should be used. As discussed in Chapters 2 and 3, the normal pitch-asynchronous linear prediction analysis can result in an error of 10% for the first formant frequency of high-pitched voices (women and children). This amount of error usually causes a change in the quality of synthesized speech. The close-phase analysis can reduce this error. Another reason for using the closed-phase analysis is that the formant bandwidth increase during the glottal opening period is accounted for by the source-tract interaction model. A closed-phase analysis algorithm can be designed using the electroglottograph (EGG) signal to identify the glottis closed point. This algorithm is currently under intensive study by a fellow student, A. Krishnamurthy [43]. He also developed an automatic glottal inverse filtering algorithm using a closed-phase analysis algorithm.

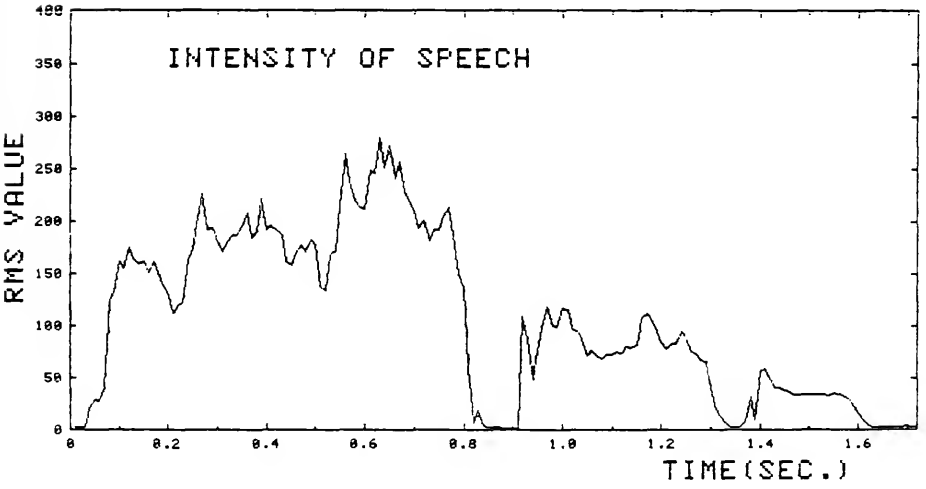
3) A better algorithm for analyzing the formant/antiformant frequencies of nasalized sound should be developed. As discussed in Chapter 3, a major problem in speech synthesis is synthesizing nasalized sounds. This is due to the difficulty in modeling the transfer function of a pole-zero system using linear analysis. The algorithm discussed in Appendix B for detecting the nasalized sound

Figure 62. Illustration of speech segmentation scheme: peaks in the intensity contour (marked *) can be used as anchor points, and valleys in the intensity or pitch period contour can be used as section boundary points.

Figure 63. The intensity (RMS value) of the speech signal.



WE WERE AWAY A YEAR AGO



may be useful in dealing with the modeling of nasalized sounds, because special pole-zero analysis can be applied to nasalized sounds after they have been identified. In particular, we need to estimate the antiformant frequency in the first formant region only, according to the model of formant synthesis.

Another potential application of the proposed formant synthesis technique is to synthesize abnormal voices related to laryngeal pathologies and identifying factors contributing to the quality of abnormal voices. This is due to the ability of the formant synthesis technique to deal with the glottal excitation function directly. In applying the technique to synthesize pathological voices, however, several things should be noted.

1) Pathological voices are usually characterized by considerable jitter and shimmer [44], which means the speech can no longer be treated as a stationary signal over several pitch periods. Thus a pitch-synchronous analysis should always be used. Also, pitch-synchronous synthesis should be used to reproduce the jitter and shimmer characteristics of pathological voices.
2) Since the vocal fold vibration pattern of pathological voices may change drastically from one pitch period to another, provision should be made for the formant synthesizer to be excited by a time-varying glottal volume velocity function. This glottal volume velocity can be obtained by a pitch-synchronous inverse filtering process.

If the above pitch-synchronous analysis and synthesis can successfully reproduce the pathological voice quality, we can then proceed to identify the contributing factors of pathological voices. This can be done by isolating one or several parameters, e.g., replacing the "time-varying" glottal

volume velocity by a "constant" stylized waveform, and synthesizing the speech again. If the isolation of a parameter results in a significant improvement in the quality of speech, this parameter is then an important contributing factor to the pathological voice quality. The results of this study are useful in speech rehabilitation training programs and laryngeal pathology detection.

Finally, a question should be asked: What is the proper amount of source-tract interaction that should be incorporated in the speech synthesizer for producing natural sounding speech? This question came up in Chapter 3 when we observed that the effect of source-tract interaction seemed to be the strongest in the male's voice, medium for the female's voice, and smallest for the child's voice. This is a very qualitative statement, however. More study is needed to decide how the glottal impedance and the vocal tract impedance vary with sex and age, and how these variations affect the quality of speech.

APPENDIX A
ILLUSTRATION OF THE SPEECH ANALYSIS AND SYNTHESIS PROGRAMS

This appendix illustrates the programs of speech analysis and synthesis by going through the procedures for the sentence "We were away a year ago." The names of the programs and their functions are listed in Table A-1. The listing of the program's interactive dialog and partial results are presented without any further discussion, because they are self-explanatory.

TABLE A-1

Program	Function
INTENSE	Compute the intensity of speech
AUTOC	Estimate the pitch period and voicing information by modified autocorrelation
PITCH	Estimate the pitch period and voicing information using the cepstrum method
FORTRACK1	Estimate the formant contour using the peak-picking method
FORTRACK2	Estimate the formant contour using the root-extraction method
PLOTFORM2	Plot the formant contour
INTERPOL	Perform a 1:2 interpolation on parameter
HANDSY3	Parameter editing program for Klatt's formant synthesizer which reads input from analysis file and modifies the parameter interactively
PAREDIT1	Speech synthesis program with glottal area excitation source
PAREDIT2	Speech synthesis program with a glottal volume velocity excitation
PAREDIT	Speech synthesis program with an impulse excitation source
ALLPOLE	LP analysis program
LPCSYN	LP synthesis program

```
INTENSE
INPUT FILE NAME?
SCHSTA3
ENTER THE STARTING POINT 0
UPDATE RATE?(POINTS) 100
TOTAL POINTS OF THE DATA? 16800
      1          6.431
      2          9.512
      3          7.267
      4         26.553
      5         29.414
      6         28.828
      7         32.280
      8         41.899
      9         42.443
     10         44.207
     11         43.843
     12         44.829
     13         44.307
     14         44.062
     15         44.164
     16         43.590
     17         44.147
     18         43.587
     19         42.797
     20         42.444
     21         40.928
     22         41.564
     23         41.713
     24         44.232
     25         44.810
     26         46.183
     27         47.101
     28         45.668
     29         45.728
     30         45.103
     31         44.662
     32         45.120
     33         45.411
     34         45.437
```

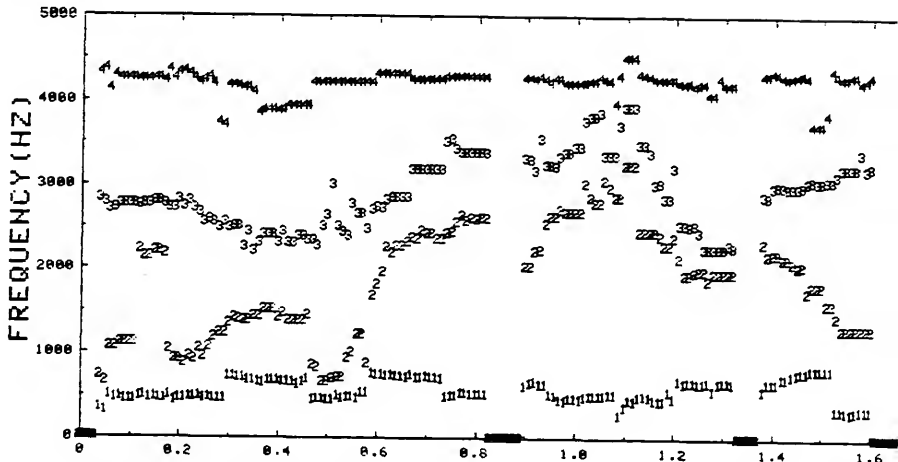
```
AUTOC
INPUT DATA FILE NAME?
SCHSTA3
OUTPUT FILE NAME FOR PITCH INFORMATION
SCHSTA3.P
THIS PROGRAM COMPUTES THE PITCH INFORMATION OF EVERY
30 MSEC. OF SPEECH, WITH ADJACENT SECTION OVERLAPPED
BY 20 MSEC.

WHERE IS THE STARTING POINTS? 0
HOW MANY POINTS TO BE PROCESSED? 16800
WHAT IS THE THRESHOLD LEVEL FOR SPEECH? 20
THRESHOLD LEVEL FOR VOICING? 0.3
SUM=      0.952469E  1
       1      -50     0.000000E  0     0.300000E  0
SUM=      0.210113E  2
    0.000000E  0        0     0.300000E  0       91
       2      -50     0.000000E  0     0.300000E  0
SUM=      0.311381E  2
    0.250000E  0        0     0.300000E  0        6
       3      -50     0.250000E  0     0.300000E  0
SUM=      0.318232E  2
    0.580645E  0        0     0.300000E  0        0
       4       35     0.580645E  0     0.300000E  0
SUM=      0.361038E  2
    0.414286E  0        1     0.300000E  0      161
       5       35     0.414286E  0     0.300000E  0
SUM=      0.394531E  2
    0.413043E  0        1     0.300000E  0      159
       6      196     0.413043E  0     0.300000E  0
SUM=      0.418031E  2
    0.483607E  0        1     0.150000E  0        3
       7       37     0.483607E  0     0.150000E  0
SUM=      0.436941E  2
    0.602740E  0        1     0.150000E  0        0
       8       40     0.602740E  0     0.150000E  0
SUM=      0.442722E  2
    0.814815E  0        1     0.150000E  0        0
       9       40     0.814815E  0     0.150000E  0
SUM=      0.451193E  2
    0.821053E  0        1     0.150000E  0        0
      10       40     0.821053E  0     0.150000E  0
SUM=      0.449981E  2
```

```
PITCH
WHAT IS THE INPUT FILE NAME?
SCHSTA3
WHAT IS THE OUTPUT FILE NAME?
SCHSTA3.Q
HOW MANY SEGMENTS? 167
HOW MANY POINTS PER SEGMENT? 200
HOW MANY POINTS IN THE OVERLAP? 100
WHERE IS THE STARTING POINT? 0
THRESHOLD FOR SPEECH? 20
THRESHOLD FOR VOICING? 0.005
       1     -50     0.000000E  0     0.500000E -2
       2     -50     0.000000E  0     0.500000E -2
       3     -50     0.000000E  0     0.500000E -2
       4       0     0.758351E -2     0.500000E -2
       5       0     0.391683E -2     0.500000E -2
       6       0     0.100560E -1     0.500000E -2
       7       0     0.220328E -2     0.500000E -2
       8       0     0.731372E -2     0.500000E -2
       9       0     0.488367E -2     0.250000E -2
      10      40     0.162479E -1     0.250000E -2
      11      40     0.147817E -1     0.250000E -2
      12      40     0.253629E -1     0.250000E -2
      13      41     0.129992E -1     0.250000E -2
      14      40     0.275605E -1     0.250000E -2
      15      41     0.135250E -1     0.250000E -2
      16      41     0.301533E -1     0.250000E -2
      17      41     0.178911E -1     0.250000E -2
      18      41     0.325376E -1     0.500000E -2
      19     201     0.283644E -1     0.500000E -2
      20      42     0.147368E -1     0.250000E -2
      21      42     0.823190E -2     0.500000E -2
```

```
FORTRACK1
INPUT SPEECH FILE NAME?
SCHSTA3
FILE NAME FOR VOICING INFORMATION?
SCHSTA3.PC
FILE NAME FOR FORMANT CONTOUR?
SCHSTA3.F
WHERE IS THE STARTING FRAME? SCHSTA3\
0
NEXT ANCHOR POINT? (TYPE 0 TO INDICATE END) 900
INITIAL FORMANT LOCATIONS? 25,40,140,200
NEXT BOUNDARY POINT? 1200
IP(K)=        1       26
IP(K)=        2       56
IP(K)=        3      144
IP(K)=        4      225
      25       40      140      200
      26       56      144      225
      26       56      144      225
      26       56      144      225

IP(K)=        1       22
IP(K)=        2      145
IP(K)=        3      203
IP(K)=        4      229
      26       56      144      225
      22      145      203      229
      22    16818      145      229
      22       56      145      229

IP(K)=        1       18
IP(K)=        2       35
IP(K)=        3      109
IP(K)=        4      145
      22       56      145      229
      18       35      109      145
      18       35      145      229
      18       35      145      229

IP(K)=        1       18
IP(K)=        2       35
IP(K)=        3       95
IP(K)=        4      154
      18       35      145      229
      18       35       95      154
      18       35      154      229
      18       35      154      229
```

```
FORTRACK2
INPUT SPEECH FILE NAME?
SCHSTA3
FILE NAME FOR PITCH DATA?
SCHSTA3.PC
FILE NAME FOR FORMANT DATA?
SCHSTA3.F
FILE NAME FOR BANDWIDTH DATA?
SCHSTA3.G
WHAT IS THE ORDER OF LP ANALYSIS? 12
WHAT IS THE STARTING POINT? (NSTRT .GE. 0) 0
 WHERE IS THE STARTING FRAME? 0
NEXT ANCHOR POINT?(TYPE 0 TO INDICATE THE END) 900
INITIAL FORMANT LOCATIONS? 500,950,2800,4200
NEXT BOUNDARY POINT? 1200
IER=         0
 FORMANT         FREQUENCY        BANDWIDTH

       1      468      131
       2     1080      187
       3     2743      126
       4     4343      223
     468     1080     2743     4343
     131      187      126      223
IER=         0
 FORMANT         FREQUENCY        BANDWIDTH

       1      496      280
       2     2737       72
       3     4185      253
     496     1080     2737     4185
     280      280       72      253
IER=         0
 FORMANT         FREQUENCY        BANDWIDTH

       1      323       61
       2      671       63
       3     2814      158
       4     4438      353
     323      671     2814     4438
      61       63      158      353
IER=         0
 FORMANT         FREQUENCY        BANDWIDTH

       1      344       79
       2      728       38
       3     1896      452
       4     2869       31
       5     4388      122
     344      728     2869     4388
      79       38       31      122
```

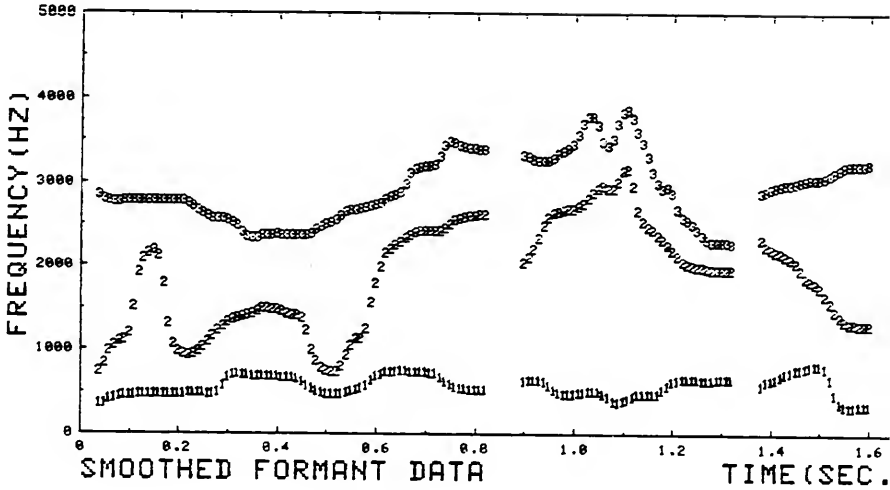
```
PLOTFORM2
 DATA FILE NAME
SCHSTA3.F
HOW MANY SETS OF DATA IN THE FILE? 172
HOW MANY FORMANTS IN A SET OF DATA? 4
MAXIMUM FREQUENCY IN THE PLOT? 4000

PLOT THE RAW DATA? 1
WANT TO CLEAR THE GRAPHIC?(YES=1,NO=0) 1
AUSE 'CHANGE THE MENU AS DESIRED'
STRIKE ANY KEY TO CONTINUE.
```




```
PAUSE 'COPY THE FORMANT DATA'
STRIKE ANY KEY TO CONTINUE.
```

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
344	728	2869	4388
357	828	2824	4401
393	978	2798	4384
422	1055	2791	4361
436	1095	2791	4343
443	1118	2792	4331
447	1200	2792	4324
454	1514	2792	4319
464	1925	2793	4313
468	2093	2796	4310
469	2159	2802	4311
469	2193	2808	4317

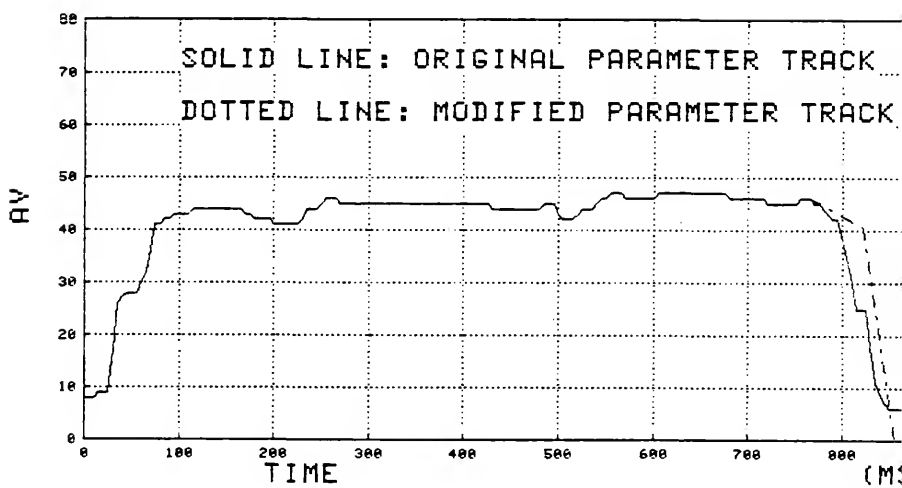
```
PAUSE 'STRIKE ANY KEY TO CONTINUE'
STRIKE ANY KEY TO CONTINUE.
WANT TO CLEAR THE GRAPHIC?(YES=1,NO=0) 1
AUSE 'CHANGE THE MENU AS DESIRED'
STRIKE ANY KEY TO CONTINUE.
```



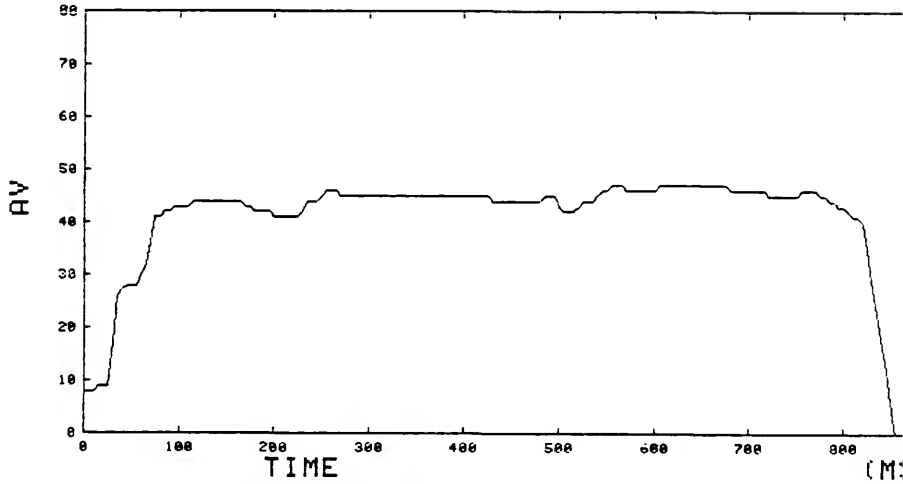

```
DO YOU WANT TO STORE THE FORMANTS IN DIFFERENT FILES?1
 FILE NAME FOR FORMANT #          1
STA3.F1
 FILE NAME FOR FORMANT #          2
STA3.F2
 FILE NAME FOR FORMANT #          3
STA3.F3
 FILE NAME FOR FORMANT #          4
STA3.F4
STOP
R


INTERPOL
INPUT FILE NAME?
SCHSTA3.PC
OUTPUT FILE NAME?
STA3.P
WHERE IS THE STARTING POINTS? 0
HOW MANY POINTS? 172
```

```
HANDSY3

CASCADE/PARALLEL FORMANT SYNTHESIZER


DOES FILE "PARAM.DOC" ALREADY EXIST(Y,N)?N

CHANGE AND/OR PRINT CONFIGURATION(Y,N)?Y

CURRENT CONFIGURATION (NAME, VAR/CON,  DEFAULT VALUE)

AV    1     0      A3    1     0      B4    0   250
AF    1     0      A4    1     0      F5    1  3750
AH    1     0      A5    1     0      B5    0   200
AVS   1     0      A6    1     0      F6    0  4900
F0    1     0      AB    1     0      B6    0  1000
F1    1   450      B1    1    50      FNP   0   250
F2    1  1450      B2    1    70      BNP   0   100
F3    1  2450      B3    1   110      BNZ   0   100
F4    1  3300      SW    0     0      BGS   0   200
FNZ   1   250      FGP   0     0      SR    0 10000
AN    0     0      BGP   0   100      NWS   0    50
A1    0     0      FGZ   0  1500      G0    0    47
A2    1     0      BGZ   0  6000      NFC   0     5

DO YOU WISH TO CHANGE WHICH PARAMETERS ARE  VARIABLE(Y,N)?Y

NAME OF PARAM TO BECOME VAR OR CON:B4
B4  IS NOW A VARIABLE

CURRENT CONFIGURATION (NAME, VAR/CON,  DEFAULT VALUE)

AV    1     0      A3    1     0      B4    1   250
AF    1     0      A4    1     0      F5    1  3750
AH    1     0      A5    1     0      B5    0   200
AVS   1     0      A6    1     0      F6    0  4900
F0    1     0      AB    1     0      B6    0  1000
F1    1   450      B1    1    50      FNP   0   250
F2    1  1450      B2    1    70      BNP   0   100
F3    1  2450      B3    1   110      BNZ   0   100
F4    1  3300      SW    0     0      BGS   0   200
FNZ   1   250      FGP   0     0      SR    0 10000
AN    0     0      BGP   0   100      NWS   0    50
A1    0     0      FGZ   0  1500      G0    0    47
A2    1     0      BGZ   0  6000      NFC   0     5

DO YOU WISH TO CHANGE WHICH PARAMETERS ARE  VARIABLE(Y,N)?N
DO YOU WISH TO CHANGE THE DEFAULT VALUE  OF A PARAMETER (Y,N)?N

CHANGE AND/OR PRINT CONFIGURATION(Y,N)?N
       1
       2
```

```
THERE ARE 21 VARIABLE PARAMETERS
PARAMETERS ARE TO BE SPECIFIED EVERY  5 MSEC
DESIRED LENGTH OF UTTERANCE IN MSEC (MAX=  980):900

DEFAULT VALUES INSERTED IN PARAMETER TRACKS

DO YOU WISH TO INSERT PARAMETER FROM A DISK FILE?Y
PARMETER TRACK MUST HAVE NUMBER OF POINTS EQUAL TO       180

NAME OF PARAMETER TRACK TO BE INSERTED? AV
WHAT IS THE FILE NAME FOR THE PARAMETER?
STA3.A
WHERE IS THE STARTING POINT? 0

DO YOU WISH TO INSERT PARAMETER FROM A DISK FILE?N

DO YOU WISH TO MODIFY A PARAMETER TRACK(Y,N)?Y
MODIFY PARAMETER TRACK BY ENTERING TIMES AND PARAMETER
VALUES. WHEN FINISHED, SIGNIFY BY ASSIGNING VALUE OF
ZERO TO T.

NAME OF PARAMETER TRACK TO BE MODIFIED? AV
```



```
DO YOU WISH TO MODIFY A PARAMETER TRACK(Y,N)?N

DO YOU WANT TO PLOT A PARAMETER TRACK(Y,N)?Y
NAME OF PARAMETER TRACK TO BE PLOTED
AV
WANT TO CLEAR THE GRAPHIC?(YES=1,NO=0) Y

PAUSE 'CHANGE THE MENU AS DESIRED'
STRIKE ANY KEY TO CONTINUE.
```




```
DO YOU WANT TO PLOT A PARAMETER TRACK(Y,N)?N
WHAT IS THE OUTPUT PARAMETER FILE NAME?
SENSTAX.DOC
PARAMETER FILE SENSTAX.DOC            SAVED

BEGIN WAVEFORM GENERATION
NNAV=        8
NDBAV=      -17
IMPULS=     0.138750E  0
NNAV=        8
NDBAV=      -17
IMPULS=     0.138750E  0
```

```
PAREDIT1
WHAT IS THE FIRST PARAMETER FILE NAME?
SENSTA1.DOC
WHAT IS THE OUTPUT SPEECH FILE NAME?
SYNSTA.1
WHAT IS THE FILENAME FOR GLOTTAL AREA FUNCTION?
STAREA.4
CONSTANT USED IN THE CALCULATION OF PS? -10
READING SYNTHESIZER CONFIGURATION FROM FILE STAREA.4

DEFAULT VALUES INSERTED IN PARAMETER TRACKS
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=        1
YGP=     0.100000E -1
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=      -49
YGP=     0.100000E -1
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=      -99
YGP=     0.100000E -1
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=     -149
YGP=     0.100000E -1
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=     -199
YGP=     0.100000E -1
```

```
PAREDIT2
WHAT IS THE FIRST PARAMETER FILE NAME?
SENSTA1.DOC
WHAT IS THE OUTPUT SPEECH FILE NAME?
SYNSTA.2
WHAT IS THE FILE NAME OF VOLUME VELOCITY FUNCTION?
STAVOL.A
READING SYNTHESIZER CONFIGURATION FROM FILE STAVOL.A

DEFAULT VALUES INSERTED IN PARAMETER TRACKS
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
INPUT=     0.000000E  0
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=        1
NPULSE1=       0
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
INPUT=     0.000000E  0
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=      -49
NPULSE1=       50
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
INPUT=     0.000000E  0
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=      -99
NPULSE1=      100
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
INPUT=     0.000000E  0
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=     -149
NPULSE1=      150
NNAV=        0
NDBAV=       -6
IMPULS=     0.500000E  0
INPUT=     0.000000E  0
IMPULS=     0.500000E  0
NPULSN=        1
NPULSE=     -199
NPULSE1=      200
```

```
PAREDIT
WHAT IS THE FIRST PARAMETER FILE NAME?
SENSTA1.DOC
WHAT IS THE OUTPUT SPEECH FILE NAME?
SYNSTA.3
READING SYNTHESIZER CONFIGURATION FROM FILE SENSTA1.DOC

DEFAULT VALUES INSERTED IN PARAMETER TRACKS
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=       0
NDBAV=      -6
IMPULS=    0.500000E  0
NNAV=      32
NDBAV=      26
IMPULS=    0.201600E  2
NNAV=      41
NDBAV=      35
IMPULS=    0.576000E  2
NNAV=      42
NDBAV=      36
IMPULS=    0.640000E  2
NNAV=      42
NDBAV=      36
IMPULS=    0.640000E  2
NNAV=      42
NDBAV=      36
IMPULS=    0.640000E  2
```

```
ALLPOLE
INPUT FILE NAME?
SCHSTA3
FILE NAME FOR PREDICTOR COEFFICIENTS?
SCHSTA3.L
FILE NAME FOR AMPLITUDE?
SCHSTA3.M
HOW MANY STRING OF DATA TO BE PROCESSED? 17200
HOW MANY POINTS PER STRING? 200
HOW MANY POINTS IN THE OVERLAP? 100
NUMBER OF PREDICTOR COEFFICIENTS? 12
PREEMPHASIZE THE DATA? (Y=1,N=0) 1
WINDOW THE DATA? 1
STARTING POINT? 0
       1
       2
       3
       4
       5
       6
       7
       8
       9
      10
      11
      12
      13
      14
      15
      16
      17
      18
      19
      20
      21
      22
      23
      24
```

```
LPCSYN
FILE NAME FOR LPC COEFFICIENTS?
SCHSTA3.L
FILE NAME FOR PITCH INFORMATION?
SCHSTA3.PD
FILE NAME FOR INTENSITY CONTOUR?
SCHSTA3.M
FILE NAME FOR OUTPUT SPEECH
SYNSTA.L
HOW MANY POINTS PER SYNTHESIS FRAME? 100
HOW MANY SYNTHESIS FRAME? 172
WHAT IS THE ORDER OF LP COEFFICIENTS? 12
     -50
     -50
     -50
     162
     117
      40
      40
      40
      40
      40
      40
      40
      40
      40
      40
      41
      41
      41
      41
      41
      41
      41
      42
```

APPENDIX B
AN ALGORITHM FOR DETECTING NASAL SOUNDS

We will discuss an algorithm we developed for detecting the existence of nasal sounds. This algorithm is useful in our formant synthesis study because it tells us when to insert the formant/antiformant pair in the vocal tract transfer function. It is also useful in speech recognition studies.

The algorithm is based on the characteristics of nasal sounds which we discussed in Chapter 3. The nasal sound characteristics may be summarized as follows:

a) An additional nasal formant/antiformant pair appears in the first formant region. In the case of a nasalized vowel, the effect is to broaden the first formant peak. For nasal consonants, the nasal antiformant almost cancels the first formant; thus, the net effect is to create a very low frequency nasal formant peak in the spectrum.

b) For the higher formant region, the formant peaks are broadened by nasalization. In the extreme case of nasal consonants, the higher peaks become obscure because of the losses introduced by the nasal tract.

The effect of nasalization can be observed in the spectrum of the nasal sounds. Figure B-1 shows the spectrum of nasal consonant /n/. We can see that there are two peaks in the first formant region due to the effect of the nasal formant/antiformant pair. Another feature is that the

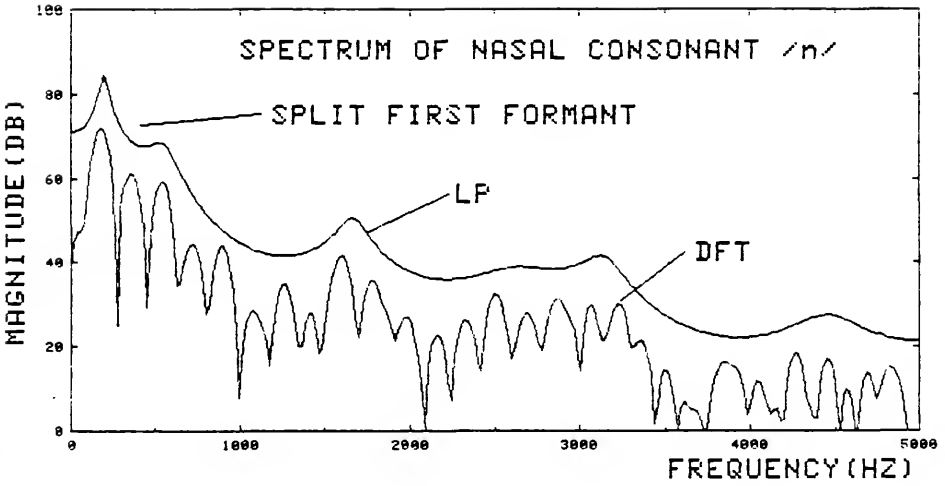


Figure B-1. Spectrum of the nasal consonant /n/.

nasal formant peak is about 25 dB higher than the second formant peak. This is due to the losses of the higher formants.

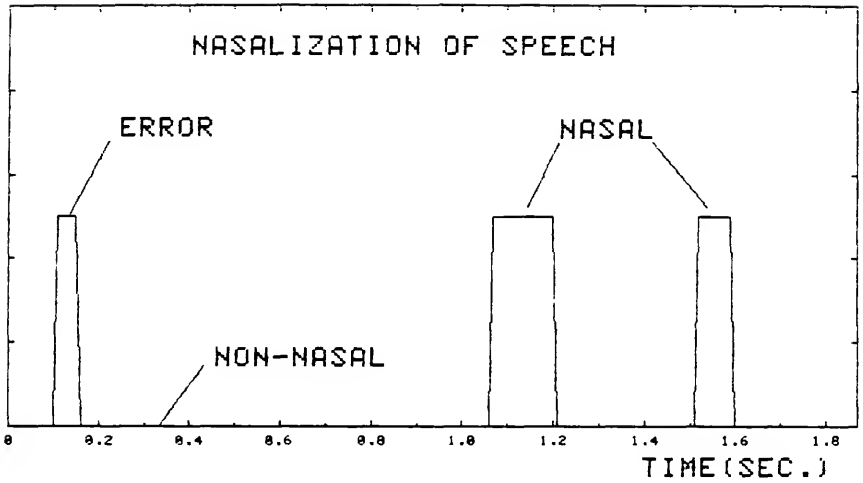
Based on the above theoretical and experimental observations, we can design an algorithm for detection of nasal sounds. The key to this algorithm is estimating the spectrum of speech using a higher order (about 18) LP analysis. This was motivated by the fact that we knew that an extra pole-zero pair exists in the vocal tract transfer function for nasalized sounds. Also, the formants are more closely spaced for nasal consonants since the nasal tract is longer than the oral tract. The following algorithm is then used to detect the nasal sounds.

1) If there are two peaks in the first formant region, then go to Step 2; if not go to Step 3.
2) Check the peaks to see if the amplitude of the first is greater than the amplitude of the second. If the answer is "yes", then the sound is a nasal consonant. Otherwise, the sound may be a nasalized vowel. To verify the latter we must check the bandwidths of both peaks.
3) If only one peak is present in the first formant region, then the algorithm checks if the amplitude of the first formant is greater than the second formant by at least 25 dB. If so, then the sound is a nasal consonant. Otherwise, if the bandwidth of the first formant is greater than 300 Hz, then the sound is a nasal vowel.
4) If none of the above hold, then the sound is a non-nasal sound.

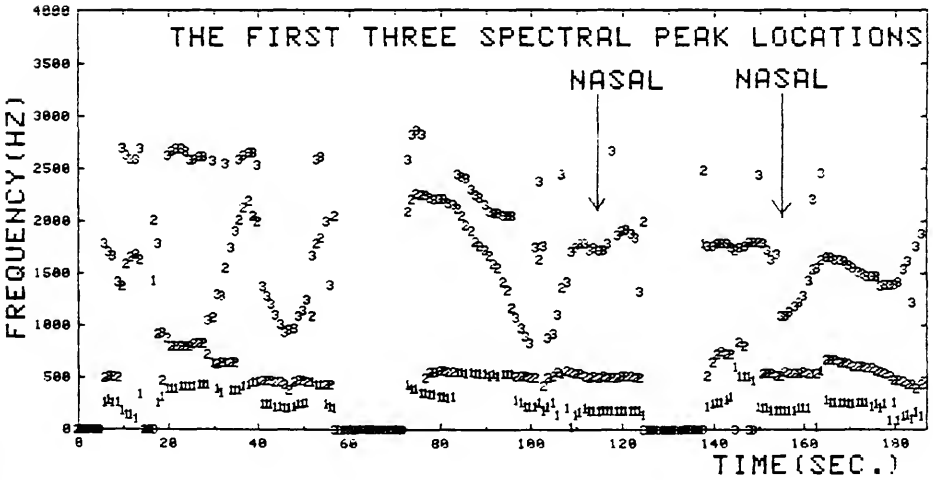
We have applied this procedure on several sentences and found that the algorithm very accurately identifies nasal consonants, but does not perform well in identifying nasal vowels. Figure B-2 shows one typical result obtained by the algorithm. The sentence is, "The boy was there

Figure B-2. The result of applying the nasal sound detection algorithm on the sentence "The boy was there when the sun rose."

Figure B-3. The first three peak locations for the sentence "The boy was there when the sun rose."



THE BOY WAS THERE WHEN THE SUN ROSE



when the sun rose." Figure B-3 shows the intermediate results, i.e., the first three peak locations. Comparing the two figures, we can see that the locations of the nasal consonants correspond to the occurrence of split first formants or a very low frequency nasal formant.

APPENDIX C
LISTENING TEST SETTING

The listening test for the student group was held in a sound treated room. The test materials are played back through an audio system consisting of a Revox tape recorder, a Sony STR-VX5 amplifier, and a pair of Infinity loudspeakers. The listeners, ten students from an audiology class, were sitting from 5 to 10 feet in front of the speakers. The loudness of the speech signal was adjusted to the most comfortable level for the listeners. The students were briefly instructed about the purpose and the nature of the listening test, and an example of the natural and synthetic speech samples were played before the test. The test sheet is shown in Figure C-1. The test is a forced-choice type. This can avoid the situation that the listener may give up due to the close resemblance of the speech signals. On the other hand, the situation of random guess can be avoided by the listener screening procedure discussed in Chapter 4.

The equipment used for the second group of listeners was the same except that a Koss PRO/4X headphone is used instead of the loudspeakers. The listening test was conducted in a normal quiet room environment because the headphone was used. All the other conditions are the same as the first listening test.

Name_____ Date_____

SESSION I	SESSION II	SESSION III
1) A B	1) A B	1) A B
2) A B	2) A B	2) A B
3) A B	3) A B	3) A B
4) A B	4) A B	4) A B
5) A B	5) A B	5) A B
6) A B	6) A B	6) A B
7) A B	7) A B	7) A B
8) A B	8) A B	8) A B
9) A B	9) A B	9) A B
10) A B	10) A B	10) A B
11) A B	11) A B	11) A B
12) A B	12) A B	12) A B
13) A B	13) A B	13) A B
14) A B	14) A B	14) A B
15) A B	15) A B	15) A B
16) A B	16) A B	16) A B
17) A B	17) A B	17) A B
18) A B	18) A B	18) A B
19) A B	19) A B	19) A B
20) A B	20) A B	20) A B
21) A B	21) A B	21) A B
22) A B	22) A B	22) A B
23) A B	23) A B	23) A B
24) A B	24) A B	24) A B

Figure C-1. Response Sheet for Listening Tests

LIST OF REFERENCES

1. J.L. Flanagan, "Voice of Men and Machines," J. Acoust. Soc. Amer. 51, 1375-1387, 1972.

2. J.L. Flanagan, K. Ishizaka, and K.L. Shipley, "Synthesis of Speech from a Dynamical Model of the Vocal Cords and Vocal Tract," Bell Sys. Tech. J. 54, 485-506, 1975.

3. C. Coker, "A Model of Articulatory Dynamics and Control," Proc. IEEE 64, 452-460, 1976.

4. B.S. Atal, J.J. Chang, M.V. Mathews, and J.W. Tukey, "Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer Sorting Technique," J. Acoust. Soc. Amer. 63, 1535-1555, 1978.

5. B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer. 50, 637-655, 1971.

6. J.N. Holmes, "Effect of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," IEEE Trans. Audio Electroacoust. AU-21, 298-305, 1973.

7. A.E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Speech," J. Acoust. Soc. Amer. 49, 583-590, 1971.

8. M.V. Mathews, J.E. Miller, and E.E. David, "Pitch Synchronous Analysis of Voiced Sound," J. Acoust. Soc. Amer. 32, 179-186, 1961.

9. D.Y. Wong & J.D. Markel, "An Excitation Function for LPC Synthesis Which Retains the Human Glottal Phase Characteristics," Proc. IEEE International Conference on Acoust., Speech, and Signal Processing, 171-174, 1979.

10. J.L. Flanagan, "Source-system Interaction in the Vocal Tract," Ann. New York Academy of Science 155, 9-17, 1968.

11. M. Nadal-Suris, Comparison of Natural and Glottal Area Waveform Synthetic Speech, Ph.D. Dissertation, University of Florida, Gainesville, Florida, 1976.

12. D.G. Childers, A. Paige, G.P. Moore, and M. Nadal-Suris, "Automation of the Measurement of Laryngeal Vibration Patterns from High Speed Film," Proc. IEEE International Conference on Acoust., Speech, and Signal Processing, 63-66, 1976.

13. G. Fant, "Glottal Source and Excitation Analysis," Speech Trans. Lab. Quarterly Progress and Status Report STL-QPSR 1/1979, 85-107, 1979.

14. M. Rothenberg, "An Interactive Model for the Voice Source," Speech Trans. Lab. Quarterly Progress and Status Report STL-QPSR 4/1981, 1-17, 1982.

15. B. Guerin, M. Mrayaty, and R. Carré, "A Voice Source Taking Account of Coupling with Supraglottal Cavities," Proc. IEEE International Conference on Acoust., Speech, and Signal Processing, 47-50, 1976.

16. D.H. Klatt, "Software for a Parallel-Serial Formant Synthesizer," J. Acoust. Soc. Amer. 67, 971-995, 1980.

17. G. Fant, Acoustic Theory of Speech Production, Mouton and Co., Gravenhage, The Netherlands, 1960.

18. J.N. Holmes, "Formant Excitation Before and After Glottal Closure," Proc. IEEE International Conference on Acoust., Speech, and Signal Processing, 39-42, 1976.

19. J.L. Flanagan, Speech Analysis, Synthesis, and Perception, Second Edition, Springer-Verlag, New York, 1972.

20. A.M. Noll, "Cepstrum Pitch Detection," J. Acoust. Soc. Amer. 41, 293-309, 1967.

21. M.M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. Audio Electroacoust. AU-16(2), 262-266, 1968.

22. R.W. Schafer and L.R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Amer. 47, 634-678, 1970.

23. J.D. Markel, "Digital Inverse Filtering, A New Tool for Formant Trajectory Estimation," IEEE Trans. Audio Electroacoust AU-20, 129-137, 1972.

24. S.S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-22, 135-141, 1974.

25. D.G. Childers, "Laryngeal Pathology Detection," CRC Critical Reviews in Bioengineering 2:4, 375-426, 1977.

26. M.G. Berouti, Estimation of Glottal Volume-velocity by the Linear Prediction Inverse-filter, Ph.D. Dissertation, University of Florida, Gainesville, Florida, 1976.

27. J.J. Yea and D.G. Childers, "Formant Synthesis: Technique to Account for Source/Tract Interaction," J. Acoust. Soc. Amer. 72, S79, 1982.

28. J.J. Yea, A.K. Krishnamurthy, J.M. Naik, G.P. Moore, and D.G. Childers, "Glottal Sensing for Speech Analysis and Synthesis," Proc. IEEE International Conference on Acoust., Speech, and Signal Processing, 1332-1335, 1983.

29. B. George, A Software Formant Cascade/Parallel Synthesizer, Master's Thesis, University of Florida, Gainesville, Florida, 1981.

30. G. Fant, "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," in: For Roman Jakobson, Mouton and Co., Gravenhage, The Netherlands, 109-120, 1956.

31. J.W. van den Berg, J.T. Zantema, and P. Doornenbal, Jr., "On the Air Resistance and the Bernoulli Effect of the Human Larynx," J. Acoust. Soc. Amer. 29(5), 626-631, 1957.

32. E. Kreyszig, Advanced Engineering Mathematics, Third Edition, John Wiley and Sons, New York, 1972.

33. J.W. van den Berg, "Direct and Indirect Determination of the Mean Subglottal Pressure," Folia Phoniat 8, 1-24, 1956.

34. P. Ladefoged and N.P. McKinney, "Loudness, Sound Pressure, and Subglottal Pressure in Speech," J. Acoust. Soc. Amer. 35, 454-460, 1963.

35. N. Isshiki, "Regulatory Mechanism of Voice Intensity Variation," J. Speech and Hearing Research 7, 17-29, 1964.

36. B. Guerin and L. J. Boe, "A Two-mass Model of the Vocal Cords: Determination of Control Parameters and Their Respective Consequences," Proc. IEEE International Conference on Acoust., Speech, and Signal Processing, 583-586, 1977.

37. "IEEE Recommended Practice for Speech Quality Measurements," IEEE Trans. on Audio and Electroacoust. AU-17(3), 225-246, 1969.

38. A.S. House, "Analog Studies of Nasal Consonants," J. Speech and Hearing Disorder 22(2), 190-204, 1957.

39. O. Fujimura, "Analysis of Nasal Consonants," J. Acoust. Soc. Amer. 34, 1865-1875, 1962.

40. T. P. Barnwell III, "Objective Fidelity Measure for Speech Coding System," J. Acoust. Soc. Amer. 65, 1658-1663, 1979.

41. E.H. Rothauser, G.E. Urbanek, and W.P. Pachl, "Isopreference Method for Speech Evaluation," J. Acoust. Soc. Amer. 44, 408-418, 1968.

42. T.V. Anathapadmanabha and G. Fant, "Calculation of True Glottal Flow and Its Components," Speech Trans. Lab. Quarterly Progress and Status Report STL-QPSR 1/1982, 1-30, 1982.

43. A. Krishnamurthy, Private Communication.

44. H. von Leden, G.P. Moore, and R. Timcke, "Laryngeal Vibrations: Measurements of the Glottic Wave, Part III: The Pathologic Larynx," A.M.A. Archives of Otolaryngology 71, 26-45, 1960.

BIOGRAPHICAL SKETCH

Jing-Jong Yea was born on November 3, 1955, in Hsin-Chu, Taiwan. In June, 1977, he was graduated from National Taiwan University in Taipei, Taiwan, with a Bachelor of Science degree in electrical engineering. In September, 1979, he attended the University of Florida, where his primary area of interest was digital signal processing and speech processing. He received his Master of Engineering degree in electrical engineering in June, 1981.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald G. Childers, Chairman
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Leon W. Couch II
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jack R. Smith
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Eugene R. Chenette
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

G. Paul Moore
Distinguished Service Professor Emeritus of Speech

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1983

Dean, College of Engineering

Dean for Graduate Studies and Research